A Methodology for Enriching a Multi-Lingual Domain Ontology using Machine Learning

Alexandros G. Valarakos†‡, Georgios Sigletos†, Vangelis Karkaletsis†,
Georgios Paliouras†, George A. Vouros‡

†Software and Knowledge Engineering Laboratory
Institute of Informatics and Telecommunications,
National Centre for Scientific Research "Demokritos"
153 10 Ag. Paraskevi, Athens, Greece
{alexv, vangelis , sigletos, paliourg}@iit.demokritos.gr

‡Department of Information and Telecommunication Systems
Engineering, School of Sciences,University of the Aegean
83200, Karlovassi, Samos, Greece
georgev@aegean.gr

Abstract. Ontologies accumulate and organize knowledge in a machine-processable and human-readable way providing a common understanding basis. Enriching a multi-lingual ontology is crucial for the success of many knowledge-based systems. We present an iterative ontology-driven methodology that enriches a multi-lingual domain ontology with new instances, exploiting machine learning techniques. The methodology is user-centered and aims to ease the task of ontology maintenance. Our first experiments show the strong dependency between the size of the initial ontology and the performance of the machine learning-based method.

1 Introduction

The World Wide Web (WWW) and the various technologies that have been emerged through the vision of the Semantic Web[1] are beneficial for many knowledge-intensive applications in various research areas. The Semantic web will consist of machine-understandable documents and data that will be easily reached and acquired. Its core technology is an artifact called '*ontology*' [2]. According to the most cited definition in the literature [3], *ontology* is an explicit specification of a domain conceptualization. It denotes and organizes entities that do exist in a domain of interest using a formal declarative language. It provides a common understanding basis through its explicitly denoted structure and vocabulary, facilitating information/knowledge dissemination and reuse. Therefore, an ontology has the potential to improve information/knowledge capturing, organization, re-use and re-finding through meticulous domain organization principles and advanced reasoning tasks.

We can categorize ontologies into the following types:

- ***Task ontology*** which organizes the problem solving structure of an existing task domain-independently [1].
- ***Domain ontology*** which organizes concepts, relations, instances that occur, as well the activities that take place, into a domain [6].
- ***Top-Level/generic/Upper-Level ontology*** which organizes generic domain-independent concepts and relations explicating important semantic distinctions [7].
- ***Application ontology*** which consists of the knowledge that models a particular application domain [6].

An ontology usually consists of two layers:

- ***The Conceptual Layer*** in which concepts, their attributes, properties and their relations are defined. In this level, the domain conceptual schema/structure is

[1] http://www.w3.org/2001/sw/

explicitly defined in a formal representation language, i.e. as a structured document e.g. *xml* or in some kind of logic e.g. *description logic*.

- ***The Instances Layer*** in which the instantiation of the conceptual structure/schema takes place. This layer comprises objects that are associated with the abstract concepts, relations and properties explicated in the conceptual layer..

Ontology learning is a research area in the context of ontology engineering that aims to reduce, as much as possible, the human effort into the labor-intensive, error-prone and time consuming process of ontology building, refinement, enrichment and maintenance by means of machine learning techniques. Moreover, *instances learning* pay special attention to populating, and enriching ontologies by extending the instances layer with new instances using machine learning methods.

Ontology enrichment is a difficult and expensive task that requires the collaboration of domain experts and knowledge engineers. The idea is to provide tools for the domain experts to discover new instances and enrich their ontologies. In this article we present an iterative domain-independent ontology-driven methodology for instances learning. The methodology takes advantage of the multi-lingual character of the domain ontologies implemented in the context of the EC-funded project CROSSMARC[2], enriching these ontologies through the acquisition of new ontology instances, from domain-specific Greek corpora. The methodology is demonstrated for the domain of vacation packages offered by travel agencies.

The overall methodology iterates through 3 stages: At the 1st stage, the domain ontology is used to semantically annotate a domain-specific corpus. At the 2nd stage, the annotated corpus is used to train a Hidden Markov Model for learning how to locate new instances. At the 3rd stage the new instances are extracted from the corpus, validated by domain experts and manually added to the domain ontology.

Section 2 describes the overall structure of CROSSMARC ontologies, whereas section 3 presents the proposed methodology for acquiring new instances. Section 4 describes the experimental settings and presents some preliminary results. Section 5 presents and compares the proposed methodology with related work. Finally, some concluding remarks are presented in section 6.

2. CROSSMARC Ontology

CROSSMARC is an EC-funded R&D project that aims to facilitate technology porting to new domains. The objective of this project is to advance technology for information extraction from Web pages in various languages employing language technology methods in conjunction to machine learning methods. CROSSMARC employs software localisation methodologies and user modelling techniques for the presentation of information extraction results according to the user's personal preferences and constraints.

CROSSMARC goal was to provide generic techniques and tools for extracting a wide range of conceivable facts in a variety of knowledge domains in four languages (English, French, Greek and Italian). The construction and maintenance of domain ontologies that are exploited by the CROSSMARC architecture was a crucial issue for the implementation of the CROSSMARC system [4]. Key roles of ontologies in the overall processing flow of CROSSMARC are the following ones:

- During *Web pages collection*, ontologies come in use as a “bag of words”. This provides a rough terminological description of the domain that helps CROSSMARC crawlers and spiders to identify interesting web pages.
- During *Information Extraction* from the collected web pages, ontologies drive the identification and classification of relevant entities in textual descriptions. Ontologies are used during fact extraction for the normalization and matching of named entities.

[2] http://www.iit.demokrtios.gr/skel/crossmarc

- During *Data Storage & Presentation,* the lexical layer (ontology layers are presented in the paragraphs that follow) of the ontology makes possible an easy rendering of an entity description from one language to another. User stereotypes include ontology attributes for information presentation stereotype preferences based on the ontology on hand.

The structure of the CROSSMARC ontologies' architecture has been designed aiming to be flexible enough in order to be
(a) applied in different domains and languages
(b) easily maintainable by modifying only the appropriate features.
For these reasons, the architecture consists of four layers:

- The *meta-conceptual layer,* which defines the ontological commitments of the CROSSMARC ontology architecture in the conceptual layer. It includes three meta-elements: Feature, Attribute and Value. These are used in the conceptual layer to assign computational semantics to elements of the ontology.
- The *conceptual layer*, which is composed by the concepts that populate the specific domain of interest. The internal representation of these concepts as well as their relations comply with the commitments been defined in the meta-conceptual layer.
- The *instances layer*, which represents domain specific individuals. Therefore, this layer instantiates each concept.
- The *lexical layer* provides the multi-lingual surface realization (lexicalization) of ontologies' concepts and instances in the four natural languages that are being supported by the project (English, Greek, French and Italian).

After a survey of existing ontology editors and tools, we decided to use Protégé-2000 [5] as the tool for ontology development and maintenance in CROSSMARC, as being one of the most well-known tools for ontology engineering that could be used among partners. We modified and improved Protégé model of representation and user-interface in order to fit CROSSMARC's user needs so as to facilitate the process of editing CROSSMARC's initial ontologies.

In the context of CROSSMARC we have implemented ontologies for three different domains: laptop offers, job offers and vacation packages offered by travel agencies. The first two provide lexicons for all 4 languages whereas the 3rd one for English and Greek. In this paper we provide examples from the 3rd ontology. The ontology consists of *part-of* relationships, which link the main concept, namely *vacation package*, with its parts (e.g. *price*, *duration*, *location*, *accommodation* etc.) Additionally, there is a *synonymy*[3] relationship for each concept of the ontology as well as the non-taxonomic "*has attribute"* relationship for each concept which links concepts and their attributes (e.g. *accommodation type*). All the above relationships are implicitly or explicitly defined in the ontology's xml schema. The ontology consists of 119 instances for the English and 116 for the Greek language.

The structure of the ontology is depicted in Fig. 1. This architecture promotes rapid customization to different languages by adding an additional lexicon for each new language and by specifying the corresponding associations between the lexicon layers, concepts and instances layers with the core ontology which includes the concepts and instances in a agreed natural language.

[3] It has to do mainly with the different surface appearance of an instance or a concept.

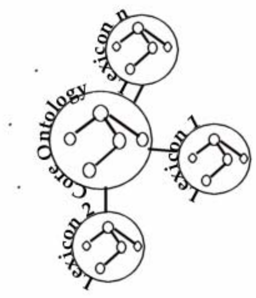

Fig. 1: CROSSMARC ontology architecture

3. Methodology

The multi-lingual nature of the CROSSMARC ontologies' architecture makes feasible the population of the ontology with instances regarding a specific surface appearance of its knowledge in the core ontology. In our case study, this work presents a method for populating the ontology with instances acquired from a domain-specific Greek corpus.

Following the methodological stages sketched in section 1, at the 1st stage, we specify the concepts of the initial ontology as well as their surface appearance. Having done this, the user is aware of the information that should validate in the validation phase. Then, we associate each concept's instances with specific Fact Types that belong to an Information Extraction xml schema. This schema specifies the types of information to be extracted by the information extraction system.

At the next stage, we use the knowledge/information coded in the ontology to annotate the training dataset [8]. An ontology instance matching is employed for creating the training examples. During this stage the ontology's instances are used to semantically annotate the corpus. The semantic annotation of the corpus is currently performed by a simple string matching technique that is biased to select the maximum spanned annotated lexical expression. The ontology-driven tagging provides training examples to the machine learning algorithm. This ontology-driven machine learning approach differs from the classical supervised methods in machine learning field as it does not use human-provided training examples but examples provided by the domain ontology using a specific error-tolerated method.

As it is usually the case, the performance of a machine learning method depends heavily on the number of training examples. For that reason, we experiment with various ontology sizes, as the "gold" ontology is unknown at runtime. We needed to investigate the size of the ontology that someone should keep up-to-date in order to maintain the ontology successfully in a periodic way.

After the training stage, the machine learning method results into a model that is capable of recognizing new instances for the concepts on which it has been trained.

In the validation stage that follows next, the user validates the new instances proposed by the machine learning method and adds them into the initial ontology. At the end of this phase the initial ontology has been enriched with new instances and the process starts again from the ontology matching (the annotation phase). The above stages are depicted in Fig. 2.

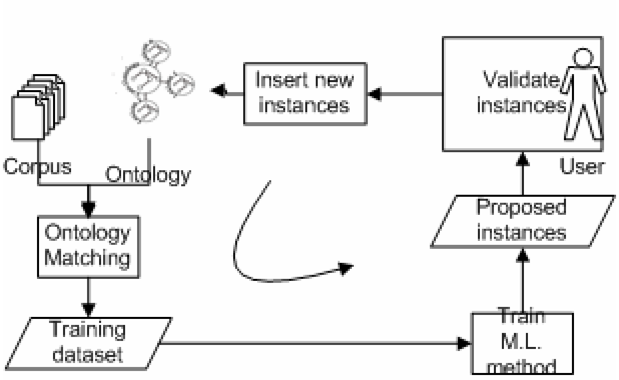

Fig. 2: Stages of the proposed methodology

4. Experiments

The corpus we used consists of 40 web pages in Greek that describe vacation packages offered by travel agencies. This corpus was semantically annotated with ontology instances so as to provide a training dataset to the machine learning method. In our case study, we trained a single Hidden Markov Model (HMM) for each of the following fact types: *name of sightseeing (Sightseeing), name of city (City), name of country (Country) and name of region (Region)* as proposed in [9] and [10]. HMM exploits the content of the ontology's instances that appear in the corpus using token-based information. Its structure is set by hand while the model parameters are estimated in a single pass over the training dataset by calculating ratios of counts (maximum likelihood estimation). At runtime, each HMM is applied to a Web page, using the *Viterbi* procedure to identify matches.

Our aim is to locate new instances that differ conceptually; we are not interested in finding synonyms (i.e. Πελοπόννησος and Μωριάς) for an instance or its potential different surface appearance (i.e. Αγία Σοφία and Αγ. Σοφία) but instances for different concepts. Therefore, we normalize the known instances found in the corpus using the ontology's first entry in the Greek lexicon for every instance. The produced model after the training of the HMM is being applied to the corpus in order to extract new instances. These are validated by the user and are manually linked to their corresponding concepts (using the Protégé-based ontology editor of CROSSMARC). The performance of this method is measured in terms of error-rate. Error-rate is defined as the ratio of erroneously extracted instances for a concept to the total number of extracted instances for this concept. Instances increase is defined as the ratio of new extracted instances for a given concept to the total number of instances for this concept.

Iteration	**Error Rate (%)**			**Instances Increase (%)**		
	Sightseeing	**City**	**Region**	**Sightseeing**	**City**	**Region**
1	64	41	28	80	27	11
2	64	49	40	0	23	20
3	64	50	38	0	36	0

Table 1: Experimental Results

Table 1 shows the results of our very first experiments. The initial ontology used has 5 instances for the concept *CITY*, 12 instances for the concept *Region* and 1 instance for the concept *Sightseeing.* We managed to enrich this tiny initial ontology following the proposed iterative methodology. The increase to the number of instances for a concept depends heavily on the initial number of instances for this concept. This is reasonable as this number influences the performance of the machine learning method that uses these instances as training examples. The error-rate is low for those concepts that are described "better". An instance is an example for a concept, thus the more the number of instances the better the description of a concept. Also, this "better" description (more training examples) of a concept helps the machine learning method to learn more successfully how to locate new examples for a particular concept.

The user sometimes found difficult to validate some unknown for him/her cities. For that reason we think that a link to the source position of the instance is needed in order the user to be helped by the context of the instance. Moreover, this link is necessary as HMMs specified sometimes the first word of the target instance i.e. “Λος” for the “Λος Άντζελες”. Using the source document in which the instance exists, the user could correct such cases increasing the performance of the methodology. Finally, a lemmatization of the instances is needed in order to increase the performance of the ontology matching phase as many instances are not encountered in the same grammatical case as they are stored in the ontology.

5. Related work

Faatz and Steinmetz [11] pose the ontology enrichment problem as a parameters optimization problem. These parameters denote how strong an initial ontology concept co-occurs with a candidate one due to a predefined rule in a text collection. A similarity measure between candidate concepts and the initial one specify the acceptance of a concept into the candidate concept group. The pairs that are above a threshold are provided to the knowledge engineer. This method acts at the conceptual level and doesn’t deal with the type of relationship holding between the candidate concept and the existing one. In contrast, our method is an iterative one and acts at the instance level manipulating a specific relationship (the ‘instance-of’) on a multi-lingual ontology.

In [12] a methodology for ontology enrichment is proposed. This occurs as the re-organization of the ontology’s concepts using a hierarchical clustering and the construction of a proposed list of topically related words for each concept. The key idea of this work is the association of each ontology concept with a collection of documents and the creation of topic signatures for each concept. Therefore, someone can manipulate the concept’s collection or its topic signatures instead of the concept. Proposed lists of words consist of words included in the corresponding topic signature. This method operates at the conceptual level of ontologies.

In [13] a cooperative methodology and a system for ontology enrichment are presented. The methodology requires the intervention of a domain expert user that validates the proposed examples (lexicalization of a relationship among concepts) which will be acquired by the system from a corpus in order to constitute the training dataset. Those examples are then fed to a wrapper induction algorithm that learns patterns able to discriminate among positive and negative examples and recognize new positive examples. This process iterates until the user feels that the system has learned to correctly identify the given relationship. In this methodology the user is employed during the learning stage, whereas in our methodology the user is used only for validating the instances that will be inserted into the ontology. Last but not least, our methodology is iterative and acts at the instances level learning on a multi-lingual ontology.

6. Concluding Remarks

We have presented a user-centered methodology that incorporates an ontology-supervised machine learning method to enrich a multi-lingual domain ontology with new instances. Our method requires only tokenization of the corpus documents from which new instances are to be acquired. Although, the proposed method has been applied on a web-based corpus it doesn’t take advantage of web pages’ special structure and features. We plan to exploit information about the web pages structure in order to investigate the effect of this additional information to the method performance.

We must also note that the results presented in this paper are only preliminary. We plan to investigate thoroughly the proposed method by extending the techniques currently used, experimenting with new machine learning techniques and applying the method in other domains.

Acknowledgements
CROSSMARC is an R&D project under the IST Programme of the European Union (IST 2000-25366). We would like to thank Maria Koutsobogera and Kanella Pouli for providing the annotated corpora as well as Georgio Kaitsa and Matina Papandreou for creating the initial ontology.

References

[1] R. Mizoguchi, K. Sinitsa and M. Ikeda, Task Ontology Design for Intelligent Educational/Training Systems, In proceedings of the ITS workshop on Architectures and Methods for Designing Cost-Effective and Reusable ITSs, Montreal, 1996.

[2] B. Chandrasekaran, J. R. Josephon and V. R. Benjamins, What are Ontologies, and Why do we need them?, IEEE Intelligent Systems, 1999

[3] T. R. Gruber, A translation approach to portable ontologies, Knowledge Acquisition, 5(2):199-220, 1993

[4] T. Pazienza, A. Stellato, M. Vindigni, A.Valarakos, V. Karkaletsis, Ontology Integration in a Multilingual e-Retail System, In Proceedings of the 2nd International Conference on Universal Access in Human-Computer Interaction, Crete, Greece, June 22-23 2003.

[5] N. F. Noy, R. W. Fergerson, & M. A. Musen. The knowledge model of Protege-2000: Combining interoperability and flexibility. 12th International Conference on Knowledge Engineering and Knowledge Management (EKAW'2000), Juan-les-Pins, France, 2000.

[6] G. van Heijst, A.Th. Schreiber, and B.J. Wielinga, Using explicit ontologies in KBs development, IJHCS, pages 183– 291, 1997.

[7] Sowa, J.F., Top-level ontological categories, In the International Journal of Human Computer Studies, 43, pp669-685, 1995.

[8] A. Valarakos, G. Sigletos, V. Karkaletsis, G. Paliouras, A Methodology for Semantically Annotating a Corpus Using a Domain Ontology and Machine Learning, to appear in RANLP, 2003

[9] Freitag, D., McCallum, A.K., Information Extraction using HMMs and shrinkage, AAAI-99 Workshop on Machine Learning for Information Extraction, pp. 31--36 (1999).

[10] Seymore, K., McCallum A.K., Rosenfeld, R., Learning hidden Markov model structure for Information Extraction, Journal of Intelligent Information Systems 8(1): 5-28, (1999).

[11] A. Faatz, R. Steinmetz, Ontology Enrichment with Texts from the WWW, In proceedings of 2nd Workshop on Semantic Web Mining, 20 August 2002, Helsinki, Finland.

[12] Agirre E., Ansa O., Martínez D., Hovy E., Enriching WordNet concepts with topic signatures, In Procceedings of the SIGLEX workshop on "WordNet and Other Lexical Resources: Applications, Extensions and Customizations", 2001.

[13] C. Brewster, F. Ciravegna and Y. Wilks: User-Centred Ontology Learning for Knowledge Management, In Proceedings of the 7th International Conference on Applications of Natural Language to Information Systems, Stockholm, June 27-28, 2002.