Bootstrapping Ontology Evolution: A Generic Approach Relying on Ontology-based Information Extraction

http://www.boemie.org/lrec2012

Monday 21 May 2012

Presenters:

Georgios Petasis, Anastasia Krithara, Alfio Ferrara, Vangelis Karkaletsis

Tutorial Programme/Overview

09:00 – 09:05 (5 min)	Welcome
09:05 – 09:30 (25 min)	Part I: Introduction to Ontology Learning Definition of Ontology Learning Existing ontology learning frameworks and systems
09:30 – 10:30 (60 min)	Part II: Semi-automated approach for ontology learning Ontology-based Information Extraction Ontology Population Ontology Enrichment Bootstrapping approach
10:30 – 11:00 (30 min)	Coffee Break
11:00 – 12:00 (60 min)	Part III: Ontology and Instance Matching Ontology matching Instance matching Application on Ontology Evolution
12:00 – 12:30 (30 min)	Part IV: Evaluating Ontology Learning Methods State-of-the-art evaluation approaches Evaluation of the bootstrapping approach
12:30 – 13:00 (30 min)	Part V: Wrap-Up and Discussion

Tutorial Description/Outline/Contents

In recent years, ontologies have become extremely popular as a means for representing machine-readable knowledge. The difficulty of extracting information from the Web, that was created mainly for visualising information, has driven to the birth of the Semantic Web, which will contain much more resources than the Web and will attach machine-readable semantic information to these resources. Realizing the difficulty of designing the grant ontology for the world, research on the Semantic Web has focused on the development of domain or task-specific ontologies, which have made their appearance in fairly large numbers.

Having provided an ontology for a specific domain, the next step is to annotate semantically related Web resources. If done manually, this process is very time-consuming and error prone. At the same time, acquiring domain knowledge for ontologies is also a resource demanding and time consuming task. Thus, the automatic or semi-automatic construction, enrichment and adaptation of ontologies, is highly desired. To this end, the evolutionary aspects of ontologies have received significant research attention during the last years, as ontology engineering has reached a certain level of maturity, considering the vast amount of contemporary methods and tools for formalizing and applying knowledge representation models.

This tutorial provides a detailed introduction to the research area of ontology evolution. After a short introduction to the problem of ontology evolution and the presentation of the current state of the art (Part I), the tutorial will present in detail the ontology learning approach that has been developed in the context of the BOEMIE EU-funded research project (www.boemie.org). The tutorial will present an ontology-based information extraction system and how this system is exploited to learn an ontology in a synergetic, semi-automated approach, employing bootstrapping (Part II). The third part of the tutorial (Part III) will focus on how internal information (encoded in instances) and external knowledge sources (i.e. other ontologies and hierarchies) can be exploited in order to enhance proposals for new concepts, through instance matching. Finally, the tutorial will conclude with the state of the art in ontology evaluation, and evaluation results of the described approach on the thematic domain of athletics (Part IV).

Part I: Introduction to Ontology Learning

An ontology “is a formal, explicit specification of a shared conceptualisation” [1], essentially a machine readable/interpretable knowledge model of a domain, in the form of concepts/classes, binary relations, axioms and rules. In other words, ontologies are meta-data schemas, providing a controlled vocabulary of concepts, each with an explicitly defined and machine process-able semantics [2]. The manual acquisition of ontologies is not an easy task, as it is a time-consuming process that requires significant resources. The purpose that motivates the ontology construction is an important factor. Ontologies build for sharing common understanding among people may not necessarily match ontologies sharing domain knowledge among software agents. Automatic or semi-automatic acquisition of ontologies can ease ontology construction, reduce costs in both time and resources, but also help in creating ontologies that better match their application [3].

Ontology learning can be defined as a set of methods and techniques that automatically extract relevant concepts, relations, axioms, and rules from a corpus, or other kinds of data, in order to form

an ontology. Ontology learning can be used to build an ontology from scratch, but also to enhance an existing ontology, with the aim to reduce the time and effort needed in the ontology development process. Ontology learning approaches can be classified into three main categories, according to the type of data used to acquire an ontology:

- Learning from unstructured data: approaches in this category learn ontologies from free text or other multimedia resources such as images, videos, audio, etc. Relying on information extraction (IE) techniques, systems in this category employ either statistical approaches [4], or natural language processing approaches [3], [5], [6], [7].
- Learning from semi-structured data: this category involves eliciting an ontology from sources that have some predefined structure, such as XML Schemas, Web pages structure, etc. Typically, approaches in this category exploit both traditional data mining [6], [8] and Web content mining techniques [9].
- Learning from structured data: acquire ontologies through the extraction of concepts and relations from knowledge contained in structured data, such as databases.

A fairly recent and detailed survey of state-of-art approaches can be found in [10].

Part II: Semi-automated approach for ontology learning

Based on our experience in the area from our involvement in several relevant projects, we consider that the task of ontology learning involves the subtasks of population, enrichment, and inconsistency resolution. Ontology population is the process of adding new instances of concepts/relations into an ontology, usually by locating the corresponding object/terms and synonyms in the corpus. Ontology enrichment is the process of extending an ontology with new concepts, relations and rules. Inconsistency resolution is responsible for remedying problems introduced by population and enrichment. In addition to these subtasks, ontology evaluation is also needed in order to measure the plausibility of the learned ontology by evaluating the usefulness of the changes. Figure 1 depicts a typical ontology learning process.

Very often, ontology learning is modelled as a bootstrapping process: an initial ontology is used as a basis for learning a new ontology, which in turn substitutes the initial one and the whole process restarts. In particular, an initial ontology is used to analyse and extract information from a corpus. The extracted information is used to evolve the ontology, and through the evolved ontology the extraction of information is improved. The bootstrapping process continues until no more information can be extracted from the corpus. Here we have to note that in every cycle the consistency of the ontology is checked and in the case of inconsistency, the changes are discarded.

The approach that will be presented in this tutorial will concentrate on ontology learning from unstructured data, involving mainly the text and image modalities, typical data that can be found in Web pages. Implemented within a bootstrapping framework, the information extraction engine employed must adapt to the evolved ontology at the various bootstrapping cycles, making the use of an ontology-based information extraction (OBIE) engine a necessity. The BOEMIE approach that will be presented employs an OBIE that relies on an ontology to a greater extent than a typical OBIE system: reasoning is not only used for inferring additional knowledge from the extracted information

and ensuring consistency, but it additionally plays a central role in the decisions taken by the extraction engine. The BOEMIE IE engine employs a machine-learning based named-entity recognition system, but substitutes all subsequent IE sub-tasks (from co-reference resolution to event detection and template element filling) with various levels of reasoning [11] (deductive and abductive) and instance matching [12]. Reasoning is also used to "fuse" information across modalities, leading to a semantic interpretation of a complete multimedia resource, such as a Web page.

Once a multimedia resource has been interpreted, the type of extracted knowledge is examined in order to be classified into four ontology evolution patterns, two of which result in ontology population (adding new instances of concepts and relations to the ontology), with the remaining resulting in ontology evolution, where new concepts, relations, axioms and rules are added in the ontology, triggering a new bootstrapping cycle. The four evolution patterns are presented in the following list [13]:

- **P1**: Single concept interpretation. The background knowledge was adequate to explain identified media objects, and the extracted information can populate the ontology.
- **P2**: Multiple concept interpretation. The background knowledge was adequate as in evolution pattern P1, but for some reason (i.e. ambiguity or missing information from the ontology) multiple (and usually contradicting) interpretations have been obtained. The extracted information must be first disambiguated (through instance matching techniques) before populating the ontology.
- **P3**: Missing concept with explained media objects. Despite the fact that the background knowledge was enough to explain media objects identified in resources, the extracted knowledge remains scattered, without a coherent representation into a real object or event. This scattered information indicates insufficient background knowledge and the need to enrich the ontology.
- **P4**: Missing concept without explained media objects. The background knowledge was not enough to explain all the information (media objects) extracted by the named-entity recognition system, suggesting a need to enrich the part of the ontology that relates to the named-entity extraction engine.

The last two patterns trigger two different types of ontology evolution, that employ clustering techniques at the level of populated instances (pattern P3) or at the level of low-level modality specific analysis (i.e. term extraction in texts or unknown object detection in images). Central to the BOEMIE ontology evolution approach are the tasks of concept learning, where new concept/relation proposals are made, and concept enhancement, where proposals are enhanced through ontology matching with external knowledge sources (i.e. other ontologies). Then, enhanced proposals, along with the supporting information for their creation, are presented to a domain expert in natural language, through a suitable interface, seeking for possible modifications and approval. The task of concept learning is guided by scenarios that try to perform various operations to the hierarchy of the ontology, like splitting an existing concept into a set of sub-concepts, proposing "similar" concepts (siblings), or proposing concepts that aggregate a set of existing concepts. These scenarios are modelled after basic tree learning operators, like node splitting, node merging and creation of new nodes, which are enough to learn any hierarchy, provided that the proper scenario is applied at any learning action during bootstrapping.

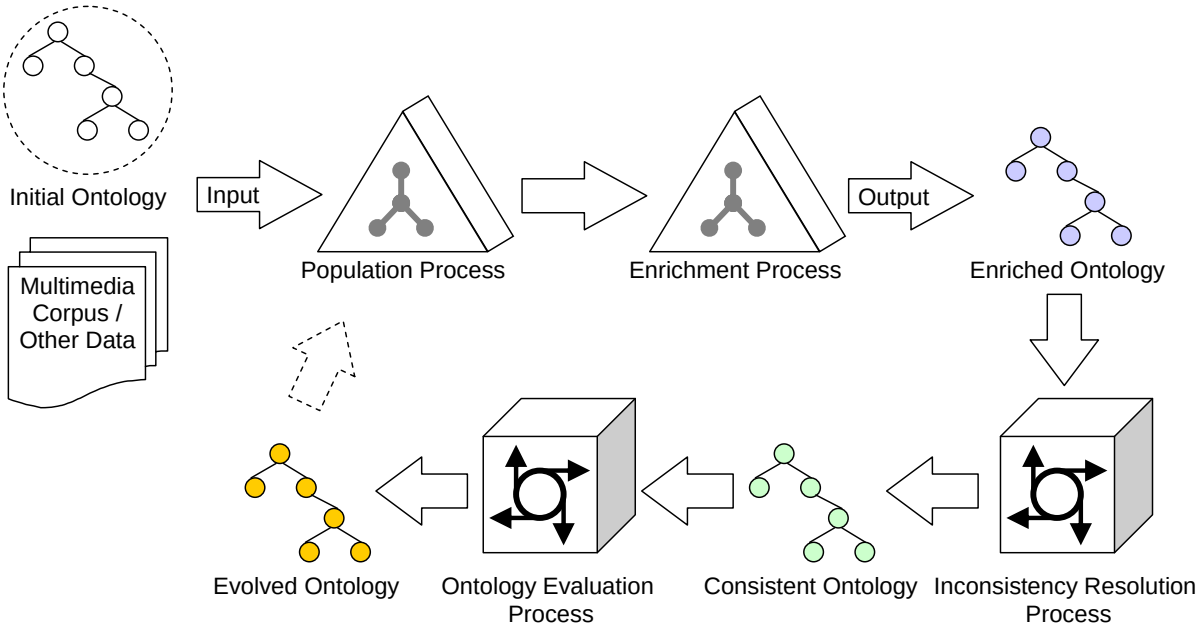

Figure 1: The process of ontology learning.

Part III: Ontology and Instance Matching

Ontology matching is the activity of developing methods and automatic or semi-automatic techniques for discovering mappings between two or more heterogeneous ontologies [14]. A mapping is a correspondence, often associated with a measure of semantic affinity or with a formal axiom, between an element of the first ontology and one or more elements of the second ontology. Analogously, instance matching is the activity of determining whether two object descriptions can be linked one to the other to represent the fact that they refer to the same real-world object in a given domain or the fact that some kind of relation holds between them [15]. In general, there are many differences between instance and ontology matching, including the fact that instance matching is more focused than ontology matching on large datasets, it is more affected by the fact that a well-defined semantics for instance identity relations is missing, and the fact that ontologies and instances are characterized by different kinds of heterogeneities and a different structure. However, in spite of these differences, both instance and ontology matching have been used for addressing the problem of merging and evolving ontologies and formal data descriptions in general.

As an example of how mappings are involved in the ontology evolution process, we refer to the work of Kondylakis et al. [16], where mappings are intended as a connection between ontologies and datasources in the context of a data integration system. In particular, the authors discuss the problem of ontology-driven data integration in a situation where ontologies are subject to changes over time and where mappings need to be evolved as the ontology changes. In this work, two approaches concerning mapping evolution are discussed: mapping composition, seen as the process of composing successive schema mappings, and mapping adaptation, seen as the activity of evolving mappings each time a change affects an ontology. Besides the problem of evolving mappings together with ontologies, which provides an interesting example of many problems related to

ontology evolution, there is the problem of using mappings as a support for the evolution of ontologies. We can identify three main roles that mappings – and related matching techniques – can play in the ontology evolution process:

- Evolving ontology by merging with other ontologies: an ontology evolves by integrating concepts and relations taken from other existing ontologies. In this case, ontology matching is useful to suggest concepts that may be integrated with the existing ones because they are similar or have the same intended meaning.
- Suggesting possible concept/instance changes by evaluating similarity among data obtained, for example, from information extraction processes. In this case, instance matching is used both for refining ontological data descriptions at the instance level and for aggregating together similar instances that could be classified under the same ontology concept.
- Measuring differences between ontology versions: instance and ontology matching are used in this case in order to evaluate the differences between different versions of the same ontology and provide a support for the evaluation of the evolution process.

In the bootstrapping ontology evolution approach, the role played by matching is twofold. On one side instance matching techniques are employed in order to identify similar instances contained in the ontology. The final goal of this task is to group together a new instance introduced in the ontology with other similar instances already contained in the ontology. Instance grouping employs clustering techniques operating on the similarity matrix returned by the instance matching task. According to the clustering results, the new instance is classified in the ontology and/or a new concept definition is suggested when needed to explain the cluster of similar instances that has been identified. On the other side, ontology matching is exploited for improving a new concept introduced in the ontology, through knowledge acquired from external sources, such as external domain ontologies or taxonomies. In particular, given the new concept, it is matched against other external concepts. As soon as similar existing concepts are identified, they can be used in order to refine the new concept definition by re-using concept properties, names, and constraints. Moreover, the new concept is matched against concepts already present in the ontology in order to suggest possible relations with the existing ontology concepts, in order to correctly collocate the new definition in the existing ontological contest.

Part IV: Evaluating Ontology Learning Methods

Evaluation in the context of ontology learning measures the quality of a learned ontology with respect to some particular criteria, in order to determine the plausibility of the learned ontology for the purposes it was built for. Approaches for evaluating learned ontologies can be distinguished into four major categories [10]:

- "Gold standard" evaluation: the learned ontology is compared to a predefined (and usually manually-constructed) "gold standard" ontology.
- Application-based evaluation: the learned ontology is used in an integrated system and is implicitly evaluated through the evaluation of the complete integrated system.
- Data-driven evaluation: the learned ontology is evaluated through comparison with a data source covering the same domain as the learned ontology.

- Human evaluation: the learned ontology is examined/evaluated by domain experts based on predefined criteria, requirements, standards, etc.

An ontology can be evaluated at different layers, such as:

- Lexical, vocabulary or data layer. The evaluation here focuses on which concepts and instances have been included in the ontology and the vocabulary used to identify them.
- Relational layer. The evaluation of this layer deals with the relations between the concepts of the ontology:
    - Hierarchy, taxonomy. An ontology almost always includes hierarchical inclusion relations between its concepts. Thus, the evaluation of these taxonomic relations is very important.
    - Semantic relations. This layer of the ontology concerns other relations besides inclusion and can be evaluated separately.
- Structure, architecture. At this layer we assess whether the design of the ontology has followed some predefined strategies and if it is possible to further develop the ontology easily.
- Philosophical layer. At this level we evaluate the ontology against highly general ontological notions, drawn from the field of philosophical ontology. Thus, we want to decide whether a property of a concept is essential for the specific concept, whether a concept is easily identified among others, etc.

Each of the approaches has different advantages and disadvantages. The majority of the evaluation approaches fall into the first category, i.e. gold standard evaluation, and the last category, i.e. evaluation by humans. These categories can also be combined and thus, they are commonly viewed as different sides of the same coin.

During the tutorial, we will present these categories in more detail.

Evaluation of the bootstrapping approach

The evaluation of the bootstrapping approach is not a trivial task, due to the need of coping with a continuously evolving ontology and with the intrinsic difficulty to univocally quantify the positive/negative impact of a change on a considered reference ontology. For this reason, we define an evaluation methodology specifically tailored to this end. In particular, what is really crucial to define, is the set of targets to evaluate during the experimentation, that have to be both measurable and meaningful at the same time. This way, we can evaluate both the capacity of the system to propose concepts/relations/rules, as well as the capability of the new concepts to explain the multimedia documents analysis results.

Bibliography

[1] R. Studer, R. Benjamins and D. Fensel, "Knowledge Engineering: Principles and Methods," *Data & Knowledge Engineering*, vol. 25, pp. 161-198, 1998.

[2] M. Hazman, S. R. El-Beltagy and A. Rafea, "Article: A Survey of Ontology Learning Approaches," *International Journal of Computer Applications*, vol. 22, pp. 36--43, May 2011.

[3] M. Sabou, C. Wroe, C. Goble and G. Mishne, "Learning domain ontologies for Web service descriptions: an experiment in bioinformatics," in *Proceedings of the 14th international conference on World Wide Web*, Chiba, Japan, 2005.

[4] D. Sanchez and A. Moreno, "Creating ontologies from Web documents," in *Recent Advances in Artificial Intelligence Research and Development*, vol. 113, IOS Press, 2004, pp. 11-18.

[5] P. Cimiano, A. Hotho and S. Staab, "Learning Concept Hierarchies from Text Corpora using Formal Concept Analysis," *Journal of Artificial Intelligence Research*, vol. 24, pp. 305-339, 2005.

[6] H. Schmid, "Probabilistic Part-of-Speech Tagging Using Decision Trees," in *Proceedings of the International Conference on New Methods in Language Processing*, Manchester, UK, 1994.

[7] P. Cimiano and J. Völker, "Text2Onto - A Framework for Ontology Learning and Data-driven Change Discovery," in *Proceedings of the 10th International Conference on Applications of Natural Language to Information Systems (NLDB)*, Alicante, Spain, 2005.

[8] N. Bennacer and L. Karoui, "A framework for retrieving conceptual knowledge from Web pages," in *Proceedings of SWAP 2005, the 2nd Italian Semantic Web Workshop, Trento, Italy, December 14-16, 2005, CEUR Workshop Proceedings*, Trento, Italy, 2005.

[9] H. Davulcu, S. Vadrevu, S. Nagarajan and I. Ramakrishnan, "OntoMiner: Bootstrapping and Populating Ontologies from Domain-Specific Web Sites," *IEEE Intelligent Systems*, vol. 18, pp. 24-33, 2003.

[10] G. Petasis, V. Karkaletsis, G. Paliouras, A. Krithara and E. Zavitsanos, "Ontology Population and Enrichment: State of the Art," vol. 6050, Springer Berlin / Heidelberg, 2011,, pp. 134--166.

[11] S. Espinosa, A. Kaya, S. Melzer and R. Möller, "On Ontology Based Abduction for Text Interpretation," in *Proceedings of 9th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2008)*, 2008.

[12] E. Iosif, G. Petasis and V. Karkaletsis, "Ontology-Based Information Extraction under a Bootstrapping Approach," in *Semi-Automatic Ontology Development: Processes and Resources*, IGI Global, 2012, pp. 1-21.

[13] S. Castano, A. Ferrara, V. Karkaletsis, A. Kaya, R. Möller, S. Montanelli and G. Petasis, "Multimedia Interpretation for Dynamic Ontology Evolution," *Journal of Logic and Computation*, vol. 19, no. 5, pp. 859-897, 2009.

[14] P. Shvaiko and J. Euzenat, "Ontology Matching: State of the Art and Future Challenges," *IEEE Transactions on Knowledge and Data Engineering*, vol. 99, 2012.

[15] A. Ferrara, A. Nikolov and F. Scharffe, "Data Linking for the Semantic Web," *International Journal on Semantic Web and Information Systems*, vol. 7, no. 3, pp. 46-76, 2011.

[16] H. Kondylakis, G. Flouris and D. Plexousakis, "Ontology and Schema Evolution in Data Integration: Review and Assessment," in *Proceedings of the Confederated International Conferences, CoopIS, DOA, IS, and ODBASE 2009 on On the Move to Meaningful Internet Systems: Part II (OTM '09)*, Vilamoura, Portugal, 2009.