Ontology engineering and knowledge extraction for crosslingual retrieval

Jantine Trapman
Utrecht University
jeandix23@yahoo.com

Paola Monachesi
Utrecht University
P.Monachesi@uu.nl

Abstract

In this paper, we show that by integrating existing NLP techniques and Semantic Web tools in a novel way, we can provide a valuable contribution to the solution of the knowledge acquisition bottleneck problem. NLP techniques to create a domain ontology on the basis of an open domain corpus have been combined with Semantic Web tools. More specifically, Watson and Prompt have been employed to enhance the kick-off ontology while Cornetto, a lexical database for Dutch, has been adopted to establish a link between the concepts and their Dutch lexicalization. The lexicalized ontology constitutes the basis for the cross-language retrieval of learning objects within the LT4eL eLearning project.

Keywords

Ontology learning, eLearning, NLP, Semantic Web, cross-lingual retrieval

1 Introduction

The aim of the Language Technology for eLearning (LT4eL)[1] project is to employ Language Technology and Semantic Web resources and tools to enhance eLearning in order to develop innovative applications for education and training [8]. One important objective is to enhance the management, distribution, search and reuse of multilingual learning material [7] .

Ontologies play a relevant role in the realization of this objective. More specifically, in the LT4eL project, the ontology mediates between the user and the learning material. The relevant concepts which are attested in the learning objects constitute the backbone of the ontology. Thus, a link is created between the learning material and its conceptualization which is represented by means of the ontology allowing for the creation of individualized learning paths. However, the most important contribution of the ontology is its role as interlingua. It facilitates access to documents in various languages since it allows for cross-lingual retrieval by mediating at the conceptual level among language specific textual realizations of the concepts.

The LT4eL project has provided a prototype which has shown the feasibility of the approach that has been validated within an eLearning context. However, in order to develop a real life application, the knowledge needs to be extracted and modeled semi-automatically. Several approaches have been proposed to this end in the Natural Language Processing as well as in the Semantic Web literature, providing a valuable contribution to the solution of this problem.

Our goal is to rely on previous results and integrate, in a novel way, existing Natural Language Processing techniques such as that proposed in [2] for ontology learning from text with recent approaches emerging from the Semantic Web community. An example of which is [12], that uses dynamically selected ontologies as background knowledge to enrich existing ontologies with new concepts. Our aim is to extend (semi-automatically) the ontology developed within the LT4eL project to new domains enabling thus the cross-lingual retrieval of new learning objects. To this end, a mapping has been carried out between the ontology and various lexicalizations. In our case, the ontology has been mapped to Cornetto [14], a Dutch lexical resource. In this paper, we report our work to extend the current ontology to a new domain (i.e. music).

The structure of the paper is as follows: in the next section, we give an overview of the LT4eL project and we discuss the role that ontologies and lexicons play in supporting the learning process. In section 3, we discuss NLP techniques for ontology learning and we focus on the approach proposed by [2], which has obvious advantages in the case of our application. Section 4, shows how the ontology developed by means of NLP techniques can be enriched further by employing tools for ontology crawling, such as Watson [1] and tools for ontology merging, such as Prompt [11]. Finally, in section 5, we discuss how the ontology is mapped to an available lexical resource which has been developed for Dutch, that is Cornetto, in order to create a lexicalized ontology. The paper ends with some conclusions.

2 The LT4eL project

One of the aims of the LT4eL project is to improve the retrieval and the usability of (multilingual) learning material within a Learning Management System in order to support the learning process.

To achieve this objective, an ontology-based search functionality has been developed which is based on the following components:

- a domain ontology in the domain of the learning objects;

[1] http://www.lt4el.eu

International Conference RANLP 2009 - Borovets, Bulgaria, pages 455–459

- a lexicon for each of the languages addressed which comprises words or phrases that are mapped to concepts attested in the ontology;
- a collection of (multilingual) learning objects annotated on the basis of the ontology.

The development of the ontology which constitutes the core of the semantic search functionality is based on domain specific corpora in the area of *computing* for the various languages addressed in the project, that is Bulgarian, Czech, Dutch, English, German, Polish, Portuguese and Romanian.

Terms have been identified in the corpora and relevant concepts have been created which constitute the backbone of the domain ontology. The domain ontology has been mapped to the DOLCE upper ontology, by means of OntoWordNet [5], which is a version of WordNet mapped to DOLCE. The current ontology contains 1002 domain concepts, 169 concepts from OntoWordNet and 105 concepts from DOLCE Ultralite.

For each language represented in the project, we have developed a lexicon on the basis of the existing ontology, following [3]. The lexicons constitute the main interface between the user's query, the ontology and the semantic search functionality which is based on the ontology.

Inline annotation of the learning material is carried out on the basis of the the ontology by means of grammars implemented in the CLaRK System.[2] The regular grammars identify the relation between the domain terms in a given language and the concepts attested in the ontology. Through the annotation, a link is created between the learning material and its conceptualization which is represented by means of the ontology.

The search engine which has been developed in the LT4eL project is based on the modules previously described. In particular, when a user types a query, the search words are looked up in the lexicons of the chosen language. If lexical entries are found in the lexicon, these are related to the concepts in the ontology. The learning objects in the desired languages are retrieved on the basis of the set of found concepts. We refer to [6] and [9] for more details.

3 Ontology learning from open domain corpora

The semantic search architecture, described in the previous section, has been developed on the basis of a manually created ontology and a corpus in the *computing* domain. However, in order for the LT4eL eLearning prototype to develop into a real life application, it is necessary to create new domain ontologies and more general lexicons. It is well known that the manual creation of an ontology is a time-consuming and expensive process. Therefore, we need to create and extend domain ontologies semi-automatically on the basis of existing resources. In this paper, we explore an approach aiming at the integration of current NLP techniques and available Semantic Web tools.

[2] http://www.bultreebank.org/clark/index.html

In this section, we discuss how NLP techniques can be employed to reach this goal. Several suggestions have been made in the literature in this respect. In particular, NLP techniques can be adopted to extract terms/concepts, definitions and relations from learning material. It is thus possible to build on existing approaches which rely mainly on statistical analysis, patterns finding and shallow linguistic parsing ([10], [4] among others for an overview).

In this paper, we focus on a methodology developed by [2], in order to create a domain ontology on the basis of an open domain corpus. The main reason to adopt this approach is that it is highly compatible with the semantic search architecture assumed in the LT4eL project because the ontology extracted through this method is based on WordNet. Recall that in the LT4eL ontology, the mapping between the domain and the upper ontology occurs via OntoWordNet.

Basili et al., propose an unsupervised technique to induce domain specific knowledge from open domain corpora, on the basis of a user query. Their algorithm exploits Latent Semantic Analysis (LSA) to extract domain terminology from a large open domain corpus, as an answer to a user query. Furthermore, Conceptual Density is employed to map the inferred terms into WordNet in order to identify domain specific subregions in it. They can be considered as lexicalized kick-off ontologies for the selected domain. The main advantages of this algorythm is that it allows for the extraction of domain ontologies from WordNet on the fly without the need for domain corpora, being thus an ideal approach for our application. It can be employed to extend our ontology to new domains, more specifically we have focussed on the the *music* domain, addressed in [2].

The relevant terms are extracted through the application of LSA on the British National Corpus. The result is a terminological lexicon consisting of 181 nouns. From this lexicon a kick-off ontology has been induced which consists of 46 classes related to each other by the is-a relation. All the (numbered) leaf nodes of the ontology carry WordNet synset IDs with them. The structure of the ontology resembles WordNet but intermediate levels between two concepts are sometimes lacking. For instance, the class *quartet* is child of *quartet* > *musical_organisation* > *group*, where > denotes the is-a relation. In WordNet however, the complete subtree for this concept is *quartet* > *musical organisation / musical group* > *organisation* > *social group*> *group*.

The approach proposed in [2] has several advantages: it can be applied at run-time to an open domain corpus; in principle no specialized content is necessary. Furthermore, it is language independent and the built-in mapping to WordNet allows for easy integration in other applications related to WordNet, as in the case of the LT4eL ontology.

The result of this approach is a kick-off ontology with a taxonomic structure that constitutes the basis for further extension. In addition, a domain lexicon is produced from the terminology extraction phase. This list of extracted terms could still support a human expert in the completion of the ontology.The advantage of this list is that it contains terms that are provided with WordNet synset IDs with them. However, en-

hancing the kick-off ontology with these terms would still be a manual task. It is thus relevant to exploit existing resources and tools developed within the Semantic Web community to assess whether it is possible to extend this kick-off ontology in a semi-automatic way.

4 Semi-automatic ontology enrichment with new concepts

The growth of the Semantic Web has influenced also the availability of freely available ontologies. Reusing such resources can save the time and effort of manual labor. We have explored two possible strategies for the extension of our kick-off ontology described in the previous section which both exploit the use of existing resources.

One relies on *crawling semantic data* by means of Watson [1], which allows for the extraction of new concepts from relevant ontologies. Watson has been preferred to other tools such as Swoogle because of the quality of the documents retrieved and the availability of relevant plug-ins. The other approach relies on *merging* the kick-off ontology with existing resources i.e. other ontologies in the music domain by using Prompt [11].

4.1 Watson

Watson is a Semantic Web application that crawls the web to find semantic documents including existing ontologies, it is available both as web interface and as Protege plug-in. We have investigated both functionalities in the task of extending our kick-off ontology. The basic assumptions behind their use are that there are available resources on-line which contain the relevant domain information. This will greatly differ from domain to domain: several ontologies are available for *Law* and *Medicine*. However, in the *Music* domain only few resources are available, more specifically:

1. Music.owl (33 classes)
2. musicontology.rdfs (83 classes)
3. music.rdf (109 classes)
4. SUMO.owl (1524 classes)

The extension of our kick-off ontology through the Watson web interface has produced an ontology with 171 classes while the one extended with the plug-in contains about 120 classes. Expanding the kick-off ontology with the web interface version of Watson is more effective than using the plug-in since more classes are identified.

A closer analysis of the resulting ontologies reveals that the one created with the web version of Watson contains a larger number of abstract classes than the one expanded with the plug-in, even though additions in the upper layer were only made to generalize over the existing classes. It contains a set of classes that are related to the digital music industry. These classes were not found with the plug-in. This is because the plug-in only matches on classes, while the Watson web interface allows the user to include classes, properties, labels, comments, local names and/or literals. It seems thus that information types other than class names include valuable clues for retrieval. This increases the chance to come across sub areas of the domain a user might have ignored otherwise. The downside is they also cause noise i.e. irrelevant documents. It should be noticed, however, that even though the plug-in version is less efficient, it has the great advantage that one can make additions to an ontology *within* the editor environment.

More generally, on the basis of our experience with Watson, we conclude that the number of resources that contribute substantially to the enhancement of an existing ontology is rather limited. The size of the relevant resources available is still quite modest, which might be the reason why some trivial classes are not found (e.g. pianist, drummer, rhythm, chord, melody). The application allows for a relatively fast and efficient extension of the ontology (i.e. from 46 classes of the kick-off ontology to the 120-170 classes of the enhanced ontology). It should be noted that crawling of new semantic data not only enhances the kick-off ontology with new classes but it also improves its original structure.

4.2 Prompt

Watson is an appropriate tool for expanding our ontology with existing resources but it does not provide options for merging ontologies which may be quicker and more efficient than adding concepts one-by-one, as in the case of Watson. Merging could be preferable if both ontologies cover the same domain but have just a partial overlap.

In order to assess whether merging would be a better way to enhance an existing ontology, we have employed Prompt, a tool for semi-automatic ontology merging and alignment [11].

We have explored its functionalities by merging two ontologies from the music domain. More specifically, we have merged our kick-off music ontology with a domain ontology from the Music Ontology Specification Group[3]. The kick-off ontology consists of 46 classes in a purely taxonomic structure. It covers concepts related to music genre, musical groups, musicians and entertainers. The latter ontology contains 92 classes: 86 primitive and 6 defined classes. It includes three group of concepts: those covering simple editorial information, a second group covering music creation workflow and a third group of concepts related to events and time.

The overlap between both ontologies is not very large, because they are both rather small and they address different topics. The result of the merge is an ontology of 103 classes. About 20 classes originate from the kick-off ontology; most of them are leaf nodes. From the other ontology, 90% is present in the resulting ontology.

Prompt calculates linguistic matches and alignment possibilities in very short time, inherited properties and subclasses can be added to the target ontology within just one step and without risking (human)

[3] http://pingthesemanticweb.com/ontology/mo/musicontology.rdfs

mistakes, similar structures are also automatically detected, a tedious and time consuming task for any human to accomplish. Moreover, the results are automatically checked and after each execution step, mappings and matches are recalculated. It should be noticed that 74% of the operations involved in the merging process were suggested by Prompt; this is quite a high number and it shows that the algorithm works properly for the task.

To conclude: both Watson and Prompt are tools that provide valuable support to the task of enriching an ontology with new concepts. However, the one-by-one additions to the ontology which Watson supports leaves the ontology builder still with a significant amount of work. Especially when the resources include a substantial number of relevant concepts and a sound hierarchical structure, a merge between them is preferred. Merging seems more efficient but is also a more complex process. The ontology engineer is faced with the challenge to discover where multiple resources can be aligned or merged. Prompt gives significant support in the merging task. But for two ontologies to be merged they have to be available off-line. The results of our investigation is that it would be desirable to integrate the crawling and merging approach since Watson and Prompt can actually complement each other: with the former the user can find suitable candidates. Those candidates can be evaluated for the representation language and size which indicate possible mismatches on the language level and for coverage, respectively. Subsequently, Prompt can be used for merging (or aligning) the resources. We will explore this integration in future research.

5 Mapping the ontology to an existing lexical resource

The ontology we have obtained by combining NLP techniques with ontology enrichment tools developed within the Semantic Web community is an ontology representing the music domain that is partly mapped to WordNet. A shortcoming of the ontologies we have used to expand our kick-off ontology, is that they lack a mapping with WordNet. This property is one of the main motivations behind the creation method of the kick-off ontology since it enables an easy mapping to an existing lexicon.

Recall that in the LT4eLproject, in order to carry out cross-language retrieval of the learning objects, we rely not only on the ontology but also on language specific lexicons which are built on the basis of the formal definitions of the concepts of the ontology. If new domain ontologies are developed, a necessary condition is that new lexicons should also be built. In the LT4eL project, these lexicons were created manually. However, another possibilty, at least for Dutch, is to employ a lexical resource recently developed, that is Cornetto [14]. One of its features makes it especially interesting for our project: it is mapped to WordNet – a feature shared also by the kick-off ontology. Mapping the lexicon to the ontology becomes thus a straightforward process.

The Cornetto database is a lexical semantic database for Dutch which contains both combinatorial and semantic information i.e. semantic relations. It consists of three linguistic layers: Lexical Units (LU) which originate from the Referentie Bestand Nederlands (RBN), a collection of synsets from the Dutch WordNet which is aligned to the English WordNet 2.0, a formal upper ontology, that is SUMO. The main goal of the Cornetto project consisted in combining and aligning the RBN and Dutch WordNet. The core of Cornetto is therefore a table of Cornetto identifiers (CIDs). This table yields 1) the relations between LUs and synsets *within* the Cornetto database, and 2) between original word senses and the synsets from RBN and Dutch WordNet respectively: each synonym from Dutch WordNet is directly related to a lexical unit from the RBN.

Cornetto is a lexicon for an open domain from which we need to filter the relevant terms from the music domain. However, WordNet has been labeled with labels from the Dewey Decimal Classification which resulted in WordNet Domains. These domain labels are also integrated in Cornetto and filtering music related terms is thus a fairly easy task. It is reported in [13] that 985 concepts from WordNet 1.6 have been assigned the music label. The number of synsets extracted from Cornetto is actually much smaller: only 111 synsets. This is because only nouns have been extracted.

We have selected the kick-off ontology enhanced with Watson for the mapping task. Two cases can be identified: the ontology contains concepts with and without a WordNet identifier.

The former case involves a rather straightforward mapping: since the ontology includes WordNet identifiers a mapping with Cornetto amounts to the retrieval of WordNet identifiers in the database. The equivalence relations between WordNet and Dutch WordNet, captured in the database, automatically supply the mapping between the concepts from the ontology and the Dutch synsets and thus ultimately with the lexical units of the RBN. We have applied this strategy in case of equivalences and near-equivalences.

A more complex situation is due to multiple EQ_NEAR_SYNONYM relations that can exist among terms from two languages through EuroWordNet's ILI. There could be a situation in which one ontology concept maps to multiple Dutch synsets or a single Dutch synset associated with several concepts through the EQ_NEAR_SYNONYM relation. An example of the latter is the concept *Quartet_Composition* associated with synset number `06610307`: *quartet:5, quartette:4*. Two Dutch synsets are near-equivalents of the English one: *kwartet:1* and *kwartet:3*. These same two Dutch synsets are also near-equivalents of the WordNet synset *quartet:2, quartette:1*, associated with the concept *Quartet_Performers* in the ontology. Hence, both synsets are mapped to two different concepts from the ontology.

After the automatic mapping, 13 of the 17 concepts with a WordNet Identifier have been assigned a mapping to 15 synsets. The fact that four concepts could not be mapped is due to the fact that not all the data was available. This leaves about 140 concepts which do not have a WordNet identifier because they originate from the enrichment of the ontology through Watson

and which should be mapped to Cornetto.

The most obvious option is to carry out a syntactic mapping between the concepts and the terms from WordNet synsets in the music domain. If a concept matches any term in a WordNet synset, it will be mapped to this synset. If such a mapping has been established, a mapping to one or more Dutch synsets can be established. At the moment, there 111 Dutch synsets in the Cornetto database which are related to 126 WordNet synsets (150 English terms). So unfortunately, because the data are sparse, there are less entries and synsets in Cornetto, than there are concepts in the ontology. Preliminary investigations show that multiword phrases and ambiguity are the most common problems for optimal automatic mapping.

6 Conclusions

We have shown that the integration of exiting NLP techniques and Semantic Web tools provide a valuable contribution to the solution of the knowledge acquisition bottleneck. The integrated approach has been tested to extend the LT4eL lexicalized domain ontology to the *music* domain. In particular, on the basis of the NLP techniques proposed by [2], we have developed a kick-off ontology consisting of 46 classes related to each other by the is-a relation. In addition, all the (numbered) leaf nodes of the ontology carry WordNet synset IDs with them.

The kick-off ontology has been enhanced with new concepts by means of two Semantic Web tools. Watson, which allows for crawling of semantic data, has allowed for an extension of the kick-off ontology (46 classes) to 120 classes (plug-in version) and to 170 classes (web interface version). While Prompt has been tested for the merging of the kick-off ontology consisting of 46 classes with a new music ontology consisting of 92 classes, resulting in a new ontology of 103 classes. The version enhanced with Watson has been mapped to Cornetto by making use of the WordNet synset IDs.

The approach sketched in this paper makes possible the cross-language retrieval of learning objects, within the LT4eL project, in new domains.

References

[1] d'Aquin, M., Gridinoc, L., Sabou, M. and Angeletou, S.: (2007), Watson: Supporting next generation semantic web applications, WWW/Internet Conference 2007.

[2] Basili, R., Gliozzo, A. and Pennacchiotti, M.: (2007), Harvesting ontologies from open domain corpora: a dynamic approach, RANLP 2007, Borovets, Bulgaria.

[3] Buitelaar, P. T. Declerck, A. Frank, S. Racioppa, M. Kiesel, M. Sintek, R. Engel, M. Romanelli, D. Sonntag, B. Loos, V. Micelli, R. Porzel, P. Cimiano (2006) LingInfo: Design and Applications of a Model for the Integration of Linguistic Information in Ontologies. In: Proc. of OntoLex06, a Workshop at LREC, Genoa, Italy, May 2006.

[4] Buitelaar, P., Cimiano, P., Magnini B., (2005) Ontology from Text: Methods, Evaluation and Applications Frontiers in Artificial Intelligence and Applications Series, Vol. 123, IOS Press.

[5] Gangemi, A., Roberto Navigli, and Paola Velardi. (2003). The OntoWordNet Project: extension and axiomatisation of conceptual relations in WordNet. International Conference on Ontologies, Databases and Applications of Semantics (ODBASE 2003), Catania, Italy.

[6] Lemnitzer L., Simov K., Osenova P., Mossel E., Monachesi P. (2007) Using a domain ontology and semantic search in an eLearning environment. In: Proceedings of The Third International Joint Conferences on Computer, Information, and Systems Sciences, and Engineering. (CISSE 2007). Springer-Verlag. Berlin Heidelberg.

[7] Lemnitzer L. ,C. Vertan, A. Killing, K. Simov, D. Evans, D. Cristea and P. Monachesi. (2007). *Improving the search for learning objects with keywords and ontologies*. In: Proceedings of the ECTEL 2007 conference. Springer Verlag.

[8] Monachesi, P., L. Lemnitzer, K. Simov. Language Technology for eLearning. Proceedings of EC-TEL 2006, in Innovative Approaches for Learning and Knowledge Sharing, LNCS 0302-9743, pp. 667-672. 2006

[9] Monachesi, P., K. Simov, E. Mossel, P. Osenova, L. Lemnitzer. What ontologies can do for eLearning. Proceedings of International Conference on Interactive Mobile and Computer Aided Learning (IMCL08).

[10] Navigli, R., Velardi P. (2004) Learning Domain Ontologies from Document Warehouses and Dedicated Websites, Computational Linguistics (30-2).

[11] Noy, N. F. and Musen, M. A.: (2000), Prompt: Algorithm and tool for automated ontology merging and alignment, AAAI/IAAI, pp. 450-455.

[12] Sabou, M., dAquin, M., Motta, E. (2006) Using the Semantic Web as Background Knowledge for Ontology Mapping. International Workshop on Ontology Matching, Athens, GA.

[13] Vossen, P., Glaser, E., van Zutphen, H. and Steenwijk, R.: 2004, Validation of meaning, Wp8.1, deliverable 8.1, Irion Technologies

[14] Vossen, P., Maks, I., Segers, R., van der Vliet, H. and van Zutphen, H.: (2008), The cornetto database: Architecture and alignment issues of combining lexical units, synsets and an ontology. Proceedings of Fourth International GlobalWordNet Conference.