Socially driven ontology enrichment for eLearning

Paola Monachesi, Thomas Markus

Utrecht Institute of Linguistics OTS
Utrecht University
P.Monachesi@uu.nl, F.T.Markus@uu.nl

Abstract

One of the objectives of the Language Technologies for Life-Long Learning (LTfLL) project, is to develop a knowledge sharing system that connects learners to resources and learners to other learners. To this end, we complement the formal knowledge represented by domain ontologies with the informal knowledge emerging from tagging. We have developed an ontology enrichment pipeline that can automatically enrich a domain ontology using data extracted from social media applications, similarity measures, DBpedia knowledge base and several heuristics. An evaluation of the resulting ontology has been carried out.

1. Introduction

Ontologies can play an important role within eLearning applications (Monachesi et al., 2008). They can guide and support the learner in the learning process since they provide a formalization of the knowledge of a domain approved by an expert. In addition, they can facilitate (multilingual) retrieval and reuse of content as well as mediate access to various sources of knowledge. Ontologies, however, might be too static since they model the knowledge of the domain at a given point in time. We still lack reliable methods to deal automatically with the conceptual dynamics of evolving domains (Hepp, 2007). In addition, ontologies might be incomplete or might not correspond to the representation of the domain knowledge available to the learner. The vocabulary of the learner (especially beginners) might be different from that of domain experts and could be more sensitive to evolving terminology or less specialized terms.

In the *Language Technology for eLearning* project (LTfLL)[1], we envisage a solution to these shortcomings by merging the dynamic knowledge provided by tagging, that is available through social media applications (i.e Delicious) with the formal knowledge provided by domain ontologies.

Similarity measures are employed to identify tags which are related to the concepts of an existing ontology while a knowledge base such as DBpedia (Auer et al., 2008) is used in order to integrate the tags with the ontology. Thus, we can include not only the expert view of a given domain, that might be shared by advanced learners, but also the view of beginners who are probably using a less specialized terminology. In addition, we are able to enrich ontologies automatically, which is an important condition for eLearning applications to be scalable.

2. State of the art

There is growing attention for ontology lifecycle management which encompasses not only the creation of an ontology, but its extension and maintenance as well. Techniques include manual methods such as special wikis for ontology modification (Ghidini et al., 2009) as well as Natural Language Processing techniques that can be exploited for ontology learning (Buitelaar et al., 2005).

Social media applications with their extended use of tags provide new possibilities for ontology learning. A good overview of the basic characteristics of social tagging systems is given in Golder and Huberman (2005). They provide an overview of several uses that tags have for bookmarking systems, such as topic identification, content type, ownership, opinion and organization. In addition, they discuss the difference between tagging and traditional classification with a taxonomy and they point out a number of problems in this respect introduced by tagging, i.e. polysemy, synonymy and basic level variation. However, a number of publications remark that social tagging is not only a flexible way of classification because no pre-defined vocabulary is needed (e.g. (Marlow et al., 2006)), but in fact a viable way to discover the shared vocabulary of a community, as discussed in Marenzi et al. (2008), which can include many new community-specific terms that are not yet present in existing lexical resources or ontologies (cf. (Cattuto et al., 2008a)).

Several works have attempted to exploit the emergent semantics of tagging systems in the context of ontology learning. For example, Specia and Motta (2007) describe an approach using tag preprocessing (morphologic similarity, exclusion of isolated tags), statistical tag clustering based on co-occurrence and relation identification (by looking up terms in online ontologies). The approach makes use of online lexical resources: in order to discover whether terms are acronyms, misspellings or variations, Google and Wikipedia are employed. Finally, Angeletou et al. (2007) builds on the work by Specia and Motta (2007) and describes an approach to find relations between social tags by looking up concepts with labels corresponding to the tags in online ontologies. A practical disadvantage that it is reported is the fact that only few tags could be directly identified in ontologies at that time.

The work presented in this paper relies on a novel combination of similar techniques. It deviates from other approaches by seamlessly integrating the tags extracted from social media applications with existing domain ontologies. It is thus possible to exploit the growing number of ontologies available as result of the Semantic Web initiative and

[1] http://www.ltfll-project.org/

enhance them with the extended vocabulary arising from social data.

3. The LTfLL project: supporting social and informal learning

The main objective of the Language Technologies for Lifelong Learning (LTfLL) project, which started in March 2008, is to create next-generation support and advice services to enhance individual and collaborative building of competences and knowledge creation in educational and organizational settings. The project makes extensive use of language technology, semantic knowledge resources and cognitive models in the services.

One of the aims of the LTfLL project is to build an infrastructure for knowledge sharing, which is the Common Semantic Framework (CSF). It allows for identification, retrieval, exchange and recommendation of relevant learning objects (LOs) and of peers. It is ontology driven allowing thus for a formalization of the knowledge arising from the various stages of the learning life-cycle. This includes a formalization of the common knowledge of the domain, in a way that can support sharing and collaboration, as well as personal and community knowledge-base construction.

As already mentioned, domain ontologies offer useful support in a learning path. In our approach, we merge the dynamic knowledge provided by users/learners through tagging with the formal knowledge provided by the domain ontologies by adding tags/concepts (or instances of concepts) and relationships between concepts in the domain ontology.

We assume that a useable collection of tags can emerge from the learners' activities within social media applications such as Delicious, YouTube or Slideshare. Popular and related tags are extracted from the available data. The existing concepts in the ontology trigger the identification of new related concepts from the tags available from the social media. Figure 1 shows a visualization of an enriched domain ontology which includes new related concepts (i.e. DOM, JQuery, Ajax, JSON).

Besides supporting self-organization and the emergence of collaborative knowledge and classification, we also aim at connecting learners to other learners. To this end, the content the learner is searching and selecting is used as a trigger to get him in touch with other users who have tagged this content or used this content before him. This is the case if the learner is a novice and needs to create his own community of people with similar interests. Alternatively, if the learner is part of an already established community based on common interests, he will need to be updated with the changes in his domain(s) of interest. The learner will focus on the learning objects that are produced by people who are relevant for the domain he studies and/or people he trusts. In this way, we add a trust dimension to the search since a learner will trust the objects produced, tagged or recommended by his own network.

In the CSF, we establish an explicit link between the network of users, tagging and the resources (cf. also (Mika, 2005)). The recommendations the system provides to the learner can be viewed as an appropriate categorization of search results. They are driven by the ontological information which includes the domain, resources and learning context. It is the domain ontology that provides the necessary formalization needed to structure the heterogeneous data.

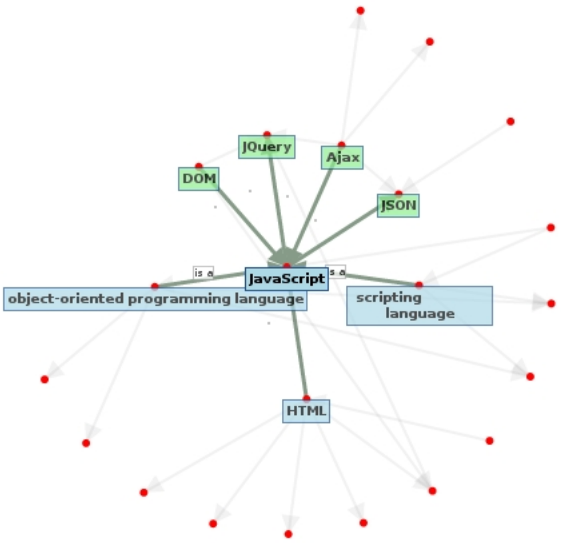

Figure 1: Automatically enriched ontology fragment. New concepts that have been added by the enrichment process are green. Existing concepts from the domain ontology are blue.

4. Enhancing ontologies with social tagging

Domain ontologies created by experts can benefit from the information extracted from social media applications for their enrichment. We take, as starting point, the LT4eL domain ontology on computing that was developed in the *Language Technology for eLearning*[2] project. It contains 1002 domain concepts, 169 concepts from OntoWordNet and 105 concepts from DOLCE Ultralite[3]. The connection between tags and concepts is established by means of language-specific lexicons, where each lexicon specifies one or more lexicalizations for each concept.

Similarity measures can play a relevant role in the automatic ontology enrichment process. They can be employed to identify whether social tags that we have extracted from Delicious represent an additional lexicalization of existing concepts, (the lexicalization of) a new concept or a more specific/general concept of an existing one. Co-occurence can provide valuable input to extract taxonomic relationships between tags, as attested by Cattuto et al. (2008b). However, Sigurbjörnsson and Van Zwol (2008) points out that this measure should first be normalized by proposing two different methods: Symmetric (according to the Jaccard coefficient) and Asymmetric. Another possibility is to split the notion of co-occurrence into user co-occurrence and resource co-occurrence. The former takes the individual users into account when calculating the co-occurrence

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

[3] http://www.loa-cnr.it/DOLCE.html

scores (Cattuto et al., 2008b). In the case of resource co-occurrence, tags are said to co-occur when added to the same resource (by different users) (Cattuto et al., 2008b). Cosine similarity is also known to provide valuable input for discovering taxonomic relationships (Cattuto et al., 2008b).

Even though the application of the various similarity measures didn't allow for a straightforward automated interpretation of the data in our domain, as discussed in more details in Monachesi and Markus (2010), we have decided to use it as first step in the ontology enrichment process. Given our eLearning application, our main goal is to include information that is relevant to a learner and his peers. We therefore assess the information implicitly contained in tag collections to obtain a sense of what is relevant and what is not in a given domain. It is this information that plays an important role for learners, especially beginners. Tagging systems provide us with a domain vocabulary which is validated as common knowledge by the community that has produced it. The similarity measure selects possible lexicalizations of concepts which are both related to the existing ones in the ontology, and which are in addition, assumed to be 'socially relevant' with respect to the input lexicalisation. More specifically, we have employed the resource coocurrence measure with assymmetric normalisation in our system for efficiency reasons and wide use in the literature. The existing lexicalisations from the domain ontology are used as seed terms for generating other related terms (tags) by employing the similarity measure. A limit on the number of related terms, as generated by the similarity measure, determines the level of relevance that is required for a term to be considered important for ontology enrichment.

However, if we want to enrich the ontology with new concepts and relations derived from the new related terms as identified by the similarity measure, we still face the problem of identifying the appropriate relationships which exist between the extracted related terms and the existing domain ontology. To this end, several heuristics are employed, which rely on the use of a large background knowledge base such as DBpedia (Auer et al., 2008). DBpedia is a community effort to extract structured information from Wikipedia and to make this information accessible on the Web.

We map related terms extracted from social media to existing DBpedia resources. The underlying assumption being that DBpedia resources can be treated as concepts for our purposes. Each resource in DBpedia has various properties, including a (multi-lingual) *label*, that we consider as lexicalisations of concepts to be included in our lexicon. There are cases in which alternative page titles are attested within the *label* property, we can thus include all of them in our lexicon. By reusing the SKOS vocabulary (Miles et al., 2005), we can differentiate between a preferred lexicalisation (the head term) and additional lexicalisations (i.e. popular and alternative terms for the same concept). In the case of an ambiguous term, we can rely on DBpedia redirections, disambiguation pages and internal structure to resolve the ambiguity.

For example, we employ DBpedia to assess whether a related term can be considered a new concept or a lexicalization of an existing one. The related term is determined to either be present as a lexicalisation of some concept in the domain ontology or to exist as an alternative lexicalisation of an existing concept as discovered through DBpedia. The newly discovered lexicalisation can then be added to existing concepts in the ontology. If the related term is found to be a new concept currently not present in the domain ontology, its additional lexicalizations and possibly synonyms are identified and the new concept is added appropriately.

Some effort has been devoted to mapping other ontologies (i.e. openCyc) onto DBpedia in order to improve both their usefulness and semantic interpretability. We exploit this information to discover new taxonomic relations. To this end, we rely on the *rdf:type* assertion which is present in DBpedia resources. More specifically, the *rdf:type* assertion between a DBpedia resource and a resource from some other ontology can be used to infer that the DBpedia concept is actually a sub-concept of the object of that statement. By retrieving the lexicalisation for the super-concept, we can discover where the new concept should be placed in the target domain ontology, assuming that the super-concept is already present.

DBpedia resources are classified according to different classification schemata and one of these are categories, as extracted from Wikipedia. Wikipedia has an actively used category system which is used to group articles. These categories are also contained in other categories resulting in a complex hierarchical structure. We exploit this structure to identify possible taxonomic relations which exist between the existing concept in the domain ontology and the related term we are trying to integrate as a new concept or lexicalisation. It can be the case that the two concepts we are considering are not directly related, but indirectly through the closest shared category higher up in the hierarchy. We can automatically calculate the closest set of shared categories for two concepts. The shared category, if it exists as a concept in the target domain ontology, will be used to add the new concept at the right place in the target domain ontology. In the case that none of the shared categories is present in the target domain ontology then both the shared category and its sub-concept (the related term) are added with the appropriate taxonomic relations to the original seed concept.

To summarize with an example: given the pre-existing domain ontology concept 'XHTML', the similarity measure system generates the term 'xslt' which is attested in DBpedia as a resource (i.e. a concept) and it shares the Wikipedia category 'XML' with the 'XHTML' concept. Given that the category 'XML' is already a concept present in the domain ontology the new concept 'XSLT' can be added as a subclass of it.

The methodology proposed allows for the enrichment of an existing ontology with the vocabulary of the Community of Practice that the user is part of. More specifically, the resulting ontology integrates the socially relevant concepts within the structure of an expert view domain ontology. The ontology enrichment process can be iterated indefinitely to increase the coverage of the ontology, but doing so will degrade the high-quality structure of the original domain ontology. Extraction methods exclusively focused on deriving ontology-like structures from tag systems cannot provide such a high quality of results due to the unavailability of explicit structural information in folksonomies, which on the contrary has been made explicit in domain ontologies.

5. Evaluation

In order to evaluate our methodology, we have compared three different ontologies:

1. the LT4eL computing ontology with the related English lexicon (1200 classes);

2. the manually enriched ontology which takes the LT4eL one as basis (1336 classes and 1672 lexical entries). This is our gold standard.

3. The automatically enriched ontology, which takes the original LT4eL ontology as basis. (2016 classes and 2325 lexical entries)

A first analysis of the lexical differences between (1) and (2) shows a difference of 80 lexicalisations. The aim of our evaluation was to assess whether the automatic enrichment process would add lexicalisations (and related concepts) that overlap with the manually added lexicalizations given a similar sub-domain.

The automatically enriched ontology has been generated by considering each coocurring tag in our Delicious data set as eligible for enrichment. The Delicious dataset we have crawled contains 598379 resources, 154476 users and 221796 tags. Related tags from our delicious dataset which could not unambiguously be linked to a single DBpedia resource have not been considered for ontology enrichment.

Even though we considered every coocurring tag as eligible for use in ontology enrichment, the lexical overlap between the manually enriched ontology and the automatic one is minimal. More specifically, 69 terms which have been added manually to the LT4eL ontology are multi-word units and are not attested in Delicious. They are representative of the expert view of the domain given their level of specificity and include terms such as: NMTOKEN attribute, XML element type declaration, XML attribute list declaration. The remaining 21 terms are attested in Delicious but only 13 of them are generated by the similarity measures and are attested in DBPpedia.

Regardless of the minimal lexical overlap between the manually and the automatic enriched ontology, it is not the case that the terms added automatically are not appropriate and are misplaced in the ontology. A preliminary verification carried out by domain experts shows that the result is satisfactory both from the point of view of added classes as well as added relations. We can thus conclude that the methodology proposed allows for an appropriate enrichment process but produces a complementary vocabulary to that of a domain expert.

6. Conclusion

We have developed an ontology enrichment pipeline that can automatically enrich a domain ontology using a combination of social tagging systems, similarity measures, the DBpedia knowledge base and several heuristics. A preliminary evaluation reveals that there is minimal overlap between the ontology produced by means of a manual enrichment process carried out by an expert and our automatic enrichment process based on tags extracted from Delicious. Both ontologies are correct from a formal point of view but the latter includes the vocabulary of the community of users, while the former it includes very specialized tags provided by an expert. It is exactly this complementarity that we wanted to achieve by embedding tags into an existing ontology and that we want to exploit in eLearning applications.

7. References

S. Angeletou, M. Sabou, L. Specia, and E. Motta. 2007. Bridging the gap between folksonomies and the semantic web: An experience report. In *Workshop: Bridging the Gap between Semantic Web and Web*, volume 2.

S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. Ives. 2008. Dbpedia: A nucleus for a web of open data. *The Semantic Web*, pages 722–735.

P. Buitelaar, P. Cimiano, and B. Magnini. 2005. Ontology learning from text: methods, evaluation and applications. *Computational Linguistics*, 32(4).

C. Cattuto, D. Benz, A. Hotho, and G. Stumme. 2008a. Semantic Analysis of Tag Similarity Measures in Collaborative Tagging Systems. *Arxiv preprint arXiv:0805.2045*.

C. Cattuto, D. Benz, A. Hotho, and G. Stumme. 2008b. Semantic grounding of tag relatedness in social bookmarking systems. *The Semantic Web-ISWC 2008*, pages 615–631.

C. Ghidini, B. Kump, S. Lindstaedt, N. Mahbub, V. Pammer, M. Rospocher, and L. Serafini. 2009. Moki: The enterprise modelling wiki. *The Semantic Web: Research and Applications*, pages 831–835.

S. Golder and B.A. Huberman. 2005. The structure of collaborative tagging systems. *Arxiv preprint cs/0508082*.

M. Hepp. 2007. Possible OntologiesHow Reality Constrains the Development of Relevant Ontologies. *IEEE Internet Computing*, pages 90–96.

I. Marenzi, E. Demidova, and W. Nejdl. 2008. LearnWeb 2.0. Integrating Social Software for Lifelong Learning. In *Proceedings of World Conference on Educational Multimedia, Hypermedia and Telecommunications*, pages 1793–1802.

C. Marlow, M. Naaman, D. Boyd, and M. Davis. 2006. Position paper, tagging, taxonomy, flickr, article, toread. In *Collaborative web tagging workshop at WWW*, volume 6.

P. Mika. 2005. Ontologies are us: A unified model of social networks and semantics. *The Semantic Web–ISWC 2005*, pages 522–536.

A. Miles, B. Matthews, M. Wilson, and D. Brickley. 2005. SKOS Core: Simple knowledge organisation for the web. *Proceedings of the International Conference on Dublin Core and Metadata Applications*, 5:12–15.

P. Monachesi and T. Markus. 2010. Using social media for ontology enrichment. In *The Semantic Web: Research and Applications*. Springer. (In press).

P. Monachesi, K. Simov, E. Mossel, P. Osenova, and L. Lemnitzer. 2008. What ontologies can do for eLearning. In *Proceedings of International Conference on Interactive Mobile and Computer Aided Learning, IMCL08*.

B. Sigurbjörnsson and R. Van Zwol. 2008. Flickr tag recommendation based on collective knowledge. In *Proceeding of the 17th international conference on World Wide Web*, pages 327–336. ACM.

L. Specia and E. Motta. 2007. Integrating folksonomies with the semantic web. *The semantic web: research and applications*, pages 624–639.