Unresolved Issues in Ontology Learning - Position Paper -

Amal Zouaq[1, 2], Dragan Gasevic[2] and Marek Hatala[3],

[1] Royal Military College of Canada, [2]Athabasca University, [3] Simon Fraser University
amal.zouaq@rmc.ca, dgasevic@sfu.ca, mhatala@sfu.ca

Abstract. Despite a number of approaches to ontology learning in the last decade, there are still a number of challenges that need to be tackled by the research community. This paper describes some of these challenges and sketches some ideas that might be beneficial for solving them.

Keywords: ontology learning, Semantic Web, challenges

1 Introduction

Ontologies are a fundamental knowledge representation structure in modern Artificial Intelligence. They are also an essential component of the Semantic Web, which uses domain ontologies to conceptualize a domain through the definition of concepts, relationships, axioms and rules. However, this heavy reliance of the Semantic Web on domain ontologies also hinders its development, as building and maintaining domain ontologies is a highly error-prone and time-consuming process. Not only does the Semantic Web require domain ontologies, but it also requires semantic markup of Web content once domain ontologies are available, which is again a tedious and non-scalable task if it is done manually. To alleviate this bottleneck, the Semantic Web community has been investigating for more than a decade how to automatize ontology building and maintenance through ***ontology learning***. Various ontology learning systems like Text-To-Onto [4], Text2Onto [7], Ontolearn [3], OntoGen [5], Abraxas [8], Texcomon [2] and OntoCmaps [1] have been proposed. In general, these tools extract ontological structures from text corpora.

This paper aims at identifying the challenges facing the ontology learning tools, and opens some questions on the way these challenges might be solved.

2 What is ontology learning?

It is now widely accepted in the community that ontology learning refers to learning its constitutive components in OWL[1]: concepts (classes), taxonomy, conceptual relationships (OWL Object Property), attributes (OWL Data Type Property), axioms (De-

[1] Web Ontology Language

fined classes) and axiom schemata (disjointness, functional properties, transitive properties, etc.). However, one can notice that the majority of the approaches focus on concept and taxonomy learning [4, 7], with very few attempts to develop the other levels [1, 17]. Through our research, exploration of literature and interaction with different end-users, we have identified a number of issues that are, to our opinion, not satisfactorily resolved or dealt with in the research community.

3 Text Understanding

The issue of text understanding refers to the ambiguity and complexity of natural language and raises the question of the availability of NLP tools able to deal with this complexity. In fact, there has been considerable progress these last few years in computational syntax and semantics with the development of robust statistical syntactic parsers and wide-coverage semantic parsers [13]. These advances will certainly facilitate the understanding of texts but it still remains true that current knowledge extraction techniques are fragmentary and generally work at the sentence-level. Building wide-coverage semantic parsers would mean a broader perspective at the discourse level with the incorporation of techniques such as anaphora resolution and discourse representation structures [15]. However, to the best of our knowledge, there is no ontology learning tool which currently adopts this approach due to the complexity of the task. In fact, ontology learning tools generally rely on shallow NLP techniques and statistical methods [7]. Moreover, even with the progress in NLP-based tools for syntactic and semantic analysis, one should expect that extending the coverage of the extraction would also result in more noisy results. Dealing with this noise is another issue that we address in Section 6. Finally, semantic analysis, as practiced by the computational semantic community, adopts formal representations that can take the form of very detailed logical expressions. However, as stated by [12], purely logical approaches produce representations that are not yet robust enough to handle real text corpora. From another perspective, since current works on ontology learning rely mainly on shallower NLP or statistical methods, they fail to handle semantic phenomena such as negation and quantification and thus are unable to produce rich conceptual relations and axioms. To overcome these shortcomings, we advocate an approach to semantic analysis which takes a middle stance between such formal approaches and shallower approaches.

4 Knowledge Extraction

As previously said, the field of ontology learning theoretically covers the extraction of a number of ontological layers in increasing order of complexity. In reality, due to their reliance on shallow NLP methods, the majority of the approaches only covers the extraction of concepts and taxonomies, and generally fails to address the more complex-levels. Thus, the implementation of deeper NLP methods is a must [16]. In particular, conceptual relationships and axiom extraction seem to be lacking in the state-of-the-art, with the exception of very few works [1, 17]. In the best case, most of the available NLP approaches to ontology learning are based on regular expressions. One disadvantage of regular expressions is that they might not discover long-distance dependencies, or they might fail to appropriately extract the right knowledge from complex structures. In our previous work [1, 2], we have proposed patterns based on dependency grammars with a syntactic-semantic interface that transforms a syntactic representation into a "semantic" one. However, similarly to the majority of ontology learning approaches which rely on a fixed number of regular expressions, our pattern knowledge base was created manually, which limits its coverage. Implementing *automatic methods for pattern learning* is one challenge that should be tackled by the ontology learning community, with pattern weighting schemes that indicate the confidence or reliability of each discovered pattern. Moreover, such a learning method would provide also a way to learn *domain-dependent patterns* as well. In fact, this research is important in order to evaluate how far we can go with the domain independence paradigm, but we are also fully aware that we might hit a limit at some point. Defining this limit would be of interest to the research community and would define a clear-cut architecture with some domain-independent and domain-dependent layers.

Besides pure knowledge extraction issues, it is also of tremendous importance to start considering how ontology learning can effectively help domain experts in their work (e.g., biological data curators) [19]. In fact, current prototypes do not really allow for much interaction with the expert. Given that ontologies are a way to formalize expert knowledge, and that some fields rely heavily on very large ontologies (e.g., biomedicine), there is a need to develop an ontology learning platform which would suggest not only new concepts and relationships to the expert, but would suggest also appropriate resources (definitions, web pages, and papers) related to a given ontological item, and would exhibit active learning capabilities by considering expert input.[2]

5 Ontological Structures Labeling

As ontological structures are learned from texts, ontology learning often takes the form of learning linguistic or lexical items. This approach is motivated by the fact that domain ontologies often represent *an interface between human and machines* rather than purely logical machine-readable metadata. However, this lexical-based approach might also lead to some problems. Firstly, some domains such as the biomedical field have evolving terminologies (e.g. known genes can be renamed) [18]. Maintaining lexical ontologies in this case seems to be a huge hurdle for the domain expert. Secondly, this creates the problem of the effective label to be associated to the ontological item (e.g. stem, lemma). In the case of relationships, this problem is even harder to solve: which lemma can we assign for example to the relationship *X can be described with Y?* If we choose *"describe",* then what is the conceptual difference with the relationship *X describes Y?*

In general, an ontological element (class, relationship) is conceptually separate from its labels, which can take various forms from a language to another (Semantic Web (en), Web sémantique (fr)) and even from a domain to another. However, to

[2] Many thanks to Prof. Melissa Haendle and Prof. Carlo Torniai of Oregon Health Sciences University for fruitful discussions on the needs of the biomedical community.

keep this notion of interface between human and machines and facilitate the ontology reading for a domain expert, there is a need to identify naming conventions and standard annotations for ontological items to increase their recognition-velocity, i.e. the ability to quickly grasp the meaning of a term via its name, for domain experts [14] but also for machines. In [14], a set of annotations associated to each ontological element is proposed such as “Display name” (the name appearing in the ontology structure) or “lexical variant”. A similar standard nomenclature would allow a certain consistency in the output of ontology learning tools. In our opinion, the “Display name” should not be related to the label contrary to what is being currently done by all ontology learning tools but should be a semantic free identifier with a set of semantic annotations. This would help the management and evolution of ontologies.

6 Ontological Structures Filtering

As we already mentioned, ontology learning extracts lexical items from texts. The question is how to identify important lexical items that should be promoted as ontological structures in the domain ontology. This also raises the issue of the nature of a concept, which is here considered as a relevant/important term. For example, while building an ontology about SCORM, an eLearning standard, the term “SCORM” is certainly relevant. However, it does not admit an instance as there is no object that could be of type “SCORM”. Nevertheless, this term will be a candidate class in the majority of ontology learning systems, and this is also the approach adopted in our own work [1, 2]. Generally, a concept is considered as a nominal expression (including multi-word expressions) that is relevant to a domain. However this widely adopted definition also raises questions. For example, given the following expressions, one can wonder if they are acceptable in a domain ontology: XML representation of content organization (yes), Aggregation of content object (may be?) and Educational use of SCORM content model component (may be?). As it can be seen, it is not always easy to differentiate what is a relevant expression (concept) and what is not.

Besides this question on the nature of concepts, there is also the notion of the statistical ranking or importance of knowledge items. In general, some ontology learning tools such as OntoGen [5] do not assign any explicit score to the extracted knowledge items while others, such as Text2Onto [7], allocate some score to the extracted knowledge using traditional metrics from information retrieval such as Relative Term Frequency (RTF), TF-IDF, or Entropy. This score is used to determine the relevance of a given item but is not used to automatically filter out the extraction. However, by looking at the precision/recall results of such systems (see for example [9, 10]), which are very low, it is obvious that there is much room for improvement both at the extraction level and at the filtering level.

Another popular weighting scheme is the use of the number of hits of a search engine to calculate the probability of a given item. However, using search engines comes at the cost of a number of issues [11] generally ignored by the ontology community. For example, search engines do not stem or lemmatise the terms. Thus, all combinations of a given term should be submitted to the search engine to obtain an appropriate (if not entirely correct) web frequency. Moreover, the number of hits refers to the number of pages containing the term rather than the frequency of the term itself. For all these reasons, relying on NLP-specific resources such as Google N-gram Corpus[3] might be an interesting avenue to explore by the ontology learning community.

Finally, graph-based metrics (Betweenness, Degree, Hits, and PageRank) were also proposed to identify relevant ontological structures in our work [1]. To our knowledge, this is the sole initiative that uses these types of metrics for ontology learning. Surprisingly, these graph-based metrics outperformed standard term relevance schemes such as TF-IDF or frequency of co-occurrence in our experiments. However, these results need to be replicated on several domains and further research need to be devoted to that aspect.

7 Ontology Evaluation

One of the last but not least issues of the ontology learning community is how to handle the appropriate evaluation of the extracted ontologies due to the lack of gold standards and resources. This hinders the development of the ontology learning field and does not enable the proper evaluation of the developed tools. While we notice a number of competitions in information retrieval (e.g. TREC[4]) or information extraction (e.g. ACE[5]), such resources do not exist for ontology learning. The experience also shows that a field starts to be more mature when resources and tools can be shared and compared. Therefore, the ontology learning community would need corpora coupled with gold standards (incorporating all the constituent knowledge items of an ontology and not only glossaries and taxonomies) mimicking the content of corpora in various domains to effectively evaluate the tools. In fact, it does not seem fair for an automatic tool to compare its output to an ontology built manually by domain experts for a number of reasons:

- The ontology learning tool does not have access to the background knowledge of experts, which is one of the oldest problems in AI. An extracted ontology can only mimic or represent the content of the knowledge source. Thus comparing such an ontology with an extensive ontology built by domain experts is not satisfactory, as it does not evaluate the possibilities of the tool but rather the lack of background knowledge of the tool.
- Another challenge is related to the domain coverage of texts. Generally, even the most extensive collection of texts will not cover sufficiently a domain. Some researchers have advocated using the Web to resolve this issue (e.g. [17]), but this may also introduce more noise, hence urging the need for efficient filtering mechanisms as explained in section 6.

As a conclusion, we believe that the first challenge of an ontology learning tool should be to adequately extract meaningful information from text (with the least possible omissions of important knowledge). Thus the need of corpora and ontological gold standards is one of the most acute issues of the field.

[3] http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2006T13

[4] http://trec.nist.gov/

[5] http://projects.ldc.upenn.edu/ace/

8 Conclusion

Ontology learning is a complex process that, besides integrating deeper NLP techniques than what is currently being done in the field, is of an acute need for appropriate evaluation resources. This paper summarizes some of the current issues and open questions of the field.

Acknowledgments. This research was partly funded by the NSERC Discovery Grant Program and by the Burroughs Wellcome Fund.

References

1. Zouaq, A., Gasevic, D. and Hatala, M. (2011). Towards open ontology learning and filtering, Information Systems, Volume 36, Issue 7, Pages 1064-1081.
2. Zouaq, A. (2008). An Ontological Engineering Approach for the Acquisition and Exploitation of Knowledge in Texts, PhD Thesis, University of Montreal (in French).
3. Navigli, R. and Velardi, R.: Learning Domain Ontologies from Document Warehouses and Dedicated Web Sites. Computational Linguistics 30(2): 151-179 (2004)
4. Maedche, A. and Volz, R. (2001). The ontology extraction maintenance framework Text-To-Onto, in Proc. of the Wshp on Integrating Data Mining and Knowledge Management.
5. Fortuna, B., Grobelnik, M., and Mladenic, D. (2006). Semi-automatic Data-driven Ontology Construction System. Proc. Of the 9th Int. Multi-conf. on IS, pp. 309-318, Springer.
6. Frantzi, K.T. and Ananiadou, S. (1999). The C/NC value domain independent method for multi-word term extraction, Journal of NLP 3(6): 145-180.
7. Cimiano, P. and Völker, J. (2005). Text2Onto. NLDB 2005, pp. 227-238, Springer.
8. Brewster, C.A. (2008). Mind the gap: bridging from text to ontological knowledge, Ph.D. Thesis, University of Sheffield.
9. Brewster, C., Jupp, S., Luciano, J., Shotton D., Stevens R. and Zhang Z. (2009). Issues in learning an ontology from text. BMC Bioinformatics 10, S1.
10. Cimiano, P. Hotho, A. and Staab S. (2005). Learning concept hierarchies from text corpora using formal concept analysis. J. Artif. Int. Res. 24, 1, 305-339.
11. Adam Kilgarriff. 2007. Googleology is Bad Science. Comput. Linguist. 33, 1 (March 2007), 147-151.
12. MacCartney, B. (2009). Natural language inference. Ph.D. dissertation, Stanford Un.
13. Bos, J. (2008). Introduction to the shared task on comparing semantic representations. In Proc. of the 2008 Conf. on Semantics in Text Processing, pp. 257-261, ACL.
14. http://ontogenesis.knowledgeblog.org/948
15. Kamp, Hans and Reyle, U. 1993. From Discourse to Logic. Kluwer, Dordrecht.
16. Zouaq, A. (2010). Shallow and Deep Natural Language Processing for Ontology Learning: a Quick Overview, In Ontology Learning and Knowledge Discovery Using the Web.
17. Sanchez, D. and Moreno, A. 2008. Learning non-taxonomic relationships from web documents for domain ontology construction. Data Knowl. Eng. 64, 3, 600-623.
18. Tuason, O., L. Chen, H. Liu, J.A. Blake, and C. Friedman. Biological Nomenclature: A Source of Lexical Knowledge and Ambiguity. In: Proc. of Pac Symp Biocomput. 2004. p. 238-49.
19. Winnenburg, R, Wächter, T, Plake, C, Doms, and A, Schroeder, M. Facts from text: can text mining help to scale-up high-quality manual curation of gene products with ontologies? Brief. Bioinform. 2008;9:466–478