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