Ontology Learning from Text: Methods, Evaluation and Applications

Paul Buitelaar, Philipp Cimiano, and Bernado Magnini (editors)
(DFKI Saarbrücken, University of Karlsruhe, and ITC-irst)

Amsterdam: IOS Press (Frontiers in artificial intelligence and applications, edited by J. Breuker et al., volume 123), 2005, v+171 pp; hardbound, ISBN 1-58603-523-1, $102.00, €85.00, £59.00

Reviewed by
Christopher Brewster
University of Sheffield

This volume is a collection of extended versions of papers first presented at workshops held at the European Conference on Artificial Intelligence and the International Conference on Knowledge Engineering and Management in 2004. The editors have all made significant contributions to the field of ontology learning and have organized some of the important workshops in the area. Ontology learning has become a major area of research within the wider area of artificial intelligence and natural language processing. This is largely due to the adoption of ontologies (especially formal ontology expressed in OWL) as the standard form of knowledge representation in the Semantic Web. The vast majority of researchers approach the challenge as one of learning ontologies from texts, rather than from other possible sources. Thus, this book is timely and representative of many of the core methodologies applied.

Researchers dealing with the challenge of building ontologies from text are essentially building on the considerable array of methodologies developed in computational linguistics and natural language processing. By a judicious selection of techniques ranging from part-of-speech tagging, chunking, and parsing to clustering and IR methodologies, they attempt to deal with the three fundamental issues involved in constructing ontologies: associating terms, building hierarchies of terms and concepts, and identifying and labeling ontological relations. In many ways, ontology learning is a specialization of core computational linguistic ambitions such as automatic lexicon construction and semantic labeling of texts.

The editors begin the volume with a short overview of the ontology-learning landscape and a brief account of the papers in the volume. They present an "ontology-learning layer cake," clearly influenced by Tim Berners-Lee's Semantic Web layer cake, which starts with terms as the foundation and works up through synonyms, concepts, concept hierarchies, and relations to rules at the top. The book is then divided into three sections dealing respectively with methods, evaluation, and applications.

In the first chapter, "An information theoretic approach to taxonomy extraction for ontology learning," Pum-Mo Ryu and Key-Sun Choi extract taxonomies by using the relative specificity of a term to a domain to determine IS-A relations. They use a combination of internal NP structure ("inside information") and syntactic modifiers ("outside information") to calculate in information-theoretic terms the respective entropy of different terms. Their approach is interesting in the combination of techniques it uses. Using the MeSH thesaurus as a gold standard and Medline abstracts as a corpus, they present figures showing up to 87% precision. While their approach may be capable of confirming hierarchical relations, it is not apparent how easily it can be used to discover such relations.

In their chapter "Unsupervised text mining for the learning of DOGMA-inspired ontologies," Marie-Laurie Reinberger and Peter Spyns cluster terms by using a shallow parser and identifying some types of relations by using prepositions. Their approach is reminiscent of Grefenstette (1994) in essentially providing information concerning term association and is founded on the widely held distributional hypothesis.

Marin Kavalec and Vojtěch Svátek, in "A study on automated relation labelling in ontology learning," label relations between terms using 'concept–concept–verb' triples. These are derived using what appears to be essentially mutual information, although they term their heuristic the "above-expectation measure." They intentionally ignore the order of concepts and verbs and use stemming to collapse passive and active sentence structures. While this may be justified on sparsity grounds, much work in corpus linguistics has shown that such approaches miss important features of language.

One of the most significant papers in this collection is by Phillip Cimiano, Aleksander Pivk, Lars Schmidt-Thieme, and Steffen Staab, entitled "Learning taxonomic relations from heterogeneous sources of evidence." They integrate a number of different approaches for the learning of taxonomies proposed in the literature, including Hearst patterns, the head-matching/NP-structure technique (originally proposed by Navigli, Velardi, and Gangemi [2003]), hyponymy information from WordNet, subsumption based on corpus-based syntactic features (an extension of Grefenstette [1994]) and document-based subsumption (Sanderson and Croft 1999). A number of standard classifiers (implemented in WEKA) were used to identify the optimal combination of these different methods. The best results came from using a support-vector-machine classifier, which resulted in an *F* measure of nearly 33%. The importance of this paper lies both in the concept of integrating multiple sources and the manner of implementation. The relatively low *F* measure (in view of NLP results in general) is indicative of how great the challenge of ontology learning still remains.

The evaluation section of the book begins with "An evaluation framework for ontology enrichment" by Andreas Faatz and Ralf Steinmetz, which distinguishes ontology enrichment from ontology learning. Faatz and Steinmetz identify fundamental difficulties with the evaluation of domain- and application-specific ontologies and argue for an automated approach to evaluating any enrichment methodology. They present a number of different measures to evaluate enrichment methods, but do not show how these would work in practice or why one measure would be better than another.

Paula Velardi, Roberto Navigli, Alessandro Cucchiarelli, and Francesca Neri provide an evaluation of an actual ontology-learning system in "Evaluation of OntoLearn, a methodology for automatic learning of domain ontologies." OntoLearn is one of the more complex and sophisticated ontology-learning systems described in the literature (Navigli and Velardi 2004), and its current architecture involves five steps for which the authors provide quantitative evaluations. Qualitative evaluations are provided by automatically generating definitions for each concept by composition of WordNet and glossary definitions for the component parts of complex terms. These were evaluated by experts in two domains. Both the qualitative and quantitative approaches are highly innovative and the authors make a major contribution to ontology evaluation in this paper.

In contrast with the problem of evaluating an ontology-learning *methodology*, Robert Porzel and Rainer Malaka, in their paper "A task-based framework for ontology learning, population and evaluation," take up the challenge of evaluating specific ontologies in themselves. The authors propose a framework where a series of ontologies are

evaluated in the context of a given application. The application in their scenario is the identification of correct speech-recognition hypotheses in dialogue systems, where the correct hypotheses have been identified by hand and act as the gold standard. They use a sophisticated approach to analyzing the types of errors, identifying, for example, insertions, deletions, and substitutions needed at the level of vocabulary, IS-A relations, and other semantic relations. This is very important and innovative work, as no one before has attempted to implement an evaluation scenario where incrementally different ontologies could be evaluated.

Marta Sabou, in "Learning Web service ontologies: An automatic extraction method and its evaluation," proposes to use the documentation texts associated with Web services to extract relevant domain ontologies. She describes these texts as "sub-languages" in Grishman's sense (Grishman and Kittredge 1986), and uses standard NLP tools to develop head-matching ontological hierarchies. The paper evaluates two such ontologies for RDF(S) storage tools and bioinformatics services with respect to term extraction, suitability from an expert's perspective, and a gold standard. The main interest of the paper lies in its application to building ontologies for Semantic Web services.

The paper by Fabio Rinaldi, Elia Yuste, Gerold Schneider, Michael Hess, and David Roussel, entitled "Exploiting technical terminology for knowledge management," does not address issues concerning ontologies directly, and although full of interesting ideas fails to convey clearly what its overall intent is. The importance of term recognition in ontology learning is a recurrent theme in this volume, but the authors do not clearly link term management with the use of ontologies in knowledge management.

The final paper in the volume is by Claire Nédellec and Adeline Nazarenko, "Ontology and information extraction: A necessary symbiosis." It argues for an intimate interaction between information extraction and ontologies. As is widely recognized, the classic template of traditional (MUC-style) information extraction is a form of ontology or model of world knowledge. This paper is an excellent overview of the interaction between the two specifically with respect to the biomedical domain. In this domain, part of the challenge lies in the considerable ambiguity that exists for any given entity and also the continuous shifts in meaning due to research progress. The authors note the need for concept hierarchies for specific biological subdomains in order to obtain extraction rules at the correct level of generality. They use their previous work on the ASIUM system to construct such hierarchies and propose to improve on this by using Hearst-type pattern techniques.

This volume provides an excellent snapshot of the current state of the art in ontology learning and the related issue of ontology evaluation. It will be of interest not just to researchers involved in ontologies but also to the wider CL and NLP community who are interested to see where their component technologies are being used.

References

Grefenstette, Gregory. 1994. *Explorations in Automatic Thesaurus Discovery*. Kluwer, Dordrecht.

Grishman, Ralph and Richard Kittredge. 1986. *Analyzing Language in Restricted Domains: Sublanguage Description and Processing*. Lawrence Erlbaum Associates, Hillsdale, NJ.

Navigli, Roberto, Paola Velardi, and Aldo Gangemi. 2003. Ontology learning and its application to automated terminology translation. *IEEE Intelligent Systems*, 18(1):22–31.

Navigli, Roberto and Paula Velardi. 2004. Learning domain ontologies from document warehouses and dedicated websites. *Computational Linguistics*, 30(2):151–179.

Sanderson, Mark and W. Bruce Croft. 1999. Deriving concept hierarchies from text. In *Proceedings of the 22nd ACM SIGIR Conference*, pages 206–213. Berkeley, CA.

Christopher Brewster is a Research Associate in the Natural Language Processing Group at the University of Sheffield and is completing his Ph.D. on ontology learning from text. He has published several papers in recent years on ontology learning, evaluation, and ranking. His research interests include the Semantic Web and the relationship between language and knowledge representation. Brewster's address is Department of Computer Science, University of Sheffield, Sheffield, S1 4DP, United Kingdom; e-mail: C.Brewster@dcs.shef.ac.uk.