Book Reviews

Ontology-Based Interpretation of Natural Language

Philipp Cimiano, Christina Unger, and John McCrae
(University of Arminia Bielefeld, Germany)

Morgan & Claypool, Synthesis Lectures on Human Language Technologies,
March 2014, 178 pages, (doi:10.2200/S00561ED1V01Y201401HLT024) , $45.00

Reviewed by
Chris Biemann
TU Darmstadt, Germany

A book aiming to build a bridge between two fields that share the subject of research but do not share the same views necessarily puts itself in a difficult position: The authors have either to strike a fair balance at peril of dissatisfying both sides or nail their colors to the mast and cater mainly to one of two communities. For semantic processing of natural language with either NLP methods or Semantic Web approaches, the authors clearly favor the latter and propose a strictly ontology-driven interpretation of natural language. The main contribution of the book, driving semantic processing from the ground up by a formal domain-specific ontology, is elaborated in ten well-structured chapters spanning 143 pages of content.

Throughout the book, examples from the soccer domain excellently illustrate the main concepts in high machine-readable detail. The first chapter sets the scene with a motivating example: Humans have sufficient background knowledge to interpret a description of a soccer match and can grasp a lot of meaning beyond the literal content. For example, if team A scored two goals and team B eventually wins the match, team B scored at least three goals. To enable these and other kinds of inferences in machines, these need to be equipped with a domain ontology that formalizes domain knowledge and domain-specific reasoning, as well as with a mechanism to construct formal representations of natural language text. The key idea of this book is to place the ontology at the center of such an interpretation process: The domain and all its relevant semantic distinctions are defined in the ontology, thus the natural language semantic parser must only be aware of these. As opposed to generic tools such as Boxer (Bos 2008) that turn natural language into formal representations, driving the formalization of NL directly by its target ontology ensures that it can be directly consumed by further layers of interpretation, such as reasoners. The remainder of the first chapter gives a very short summary of the state of affairs in semantic interpretation and semantic parsing in NLP. Although this summary is sufficient to highlight what is missing for drawing inferences on top of natural language statements, it is—with a mere two pages—necessarily incomplete and omits even mainstream approaches such as frame semantic parsing or semantic role labeling. Further, the relation of the proposal to the Semantic Web is discussed, from which formats, interoperability, and description languages are leveraged. Primers on the RDF data format and on the soccer domain complete the chapter.

doi:10.1162/COLI_r_00223

© 2015 Association for Computational Linguistics

Chapter 2 defines the concept of ontology. Great care is taken to discuss the definition of ontologies and various ontology description languages such as OWL 2 DL, as well as their expressivity. In Chapter 3, linguistic formalisms for representing syntax and semantics of natural language are discussed and ways to connect these to the ontology are introduced. Although, in principle, there are many possible formalisms to serve the "linguistic side," the authors decide on Lexicalized Tree Adjoining Grammar (LTA, Schabes 1990) for syntactic processing. In the LTAG lexicon, each lexical entry links to all elementary trees (subtrees that specify the valid contexts in terms of constituents) it anchors, whereas the corresponding ontology grammar links each ontological concept to the projections of all lexical items that verbalize this concept. Thus, each concept is associated with an exhaustive enumeration of patterns it is expressed in. On the semantic level, Discourse Representation Theory (DRT; Kamp and Reyle 1993) is chosen as a formalism for semantic operations such as coreference and quantification. DRT and LTAG are subsequently paired in a representation called DUDES (Dependency-based Underspecified Discourse Representation structures; Cimiano 2009). Here, the DRT-inspired representations are again linked to ontology concepts. Alhough the choice of framework seems largely rooted in previous works of the first author, it is clearly stated that the connection to the ontology could also be realized for other syntactic (e.g., LFG, HPSG) and semantic (e.g., GLUE, MRS) frameworks.

Chapter 4 deals with the representation of the ontological lexicons, which specify the interpretation of lexical items with respect to the target ontology. The declarative LEMON lexicon model for ontologies, following closely the approach to the lexicon of Ontological Semantics (Nirenburg and Raskin 2004), is laid out and exemplified in great detail. In the fifth chapter, the authors describe how the ontology grammar for DUDES—at least, the ontology-specific parts—can be generated from the lexicon. While this might seem merely a syntactic manipulation as the patterns of concept manifestation have been encoded in the lexicon, the point is to keep the lexicon separate from DUDES to allow for the generation of other types of ontology grammars. Note that different domain ontologies generate different grammars that could yield different interpretations for the same sentence. In Chapter 6, the overall interplay of previously discussed ingredients for semantic interpretation is exemplified, and challenges with respect to structural ambiguities and underspecification are pointed out. These are supposed to be left underspecified until they can be resolved with ontological reasoning, which is the subject of Chapter 7. Subsequently, the formalization of time in the framework is declared at length in Chapter 8, building on time interval calculi from the literature. In Chapter 9, the application of question answering over structured data is discussed—not to be confused with question answering from unstructured sources in the flavor of TREC or IBM Watson. In such a system, the natural language query is translated into a SPARQL query that can be run against public endpoints. While the examples successfully illustrate the translation from DUDES representations to SPARQL queries, no qualitative or quantitative evaluation is provided despite the existence of public benchmarks.

The final chapter concludes with a very strong statement: The mainstream of computational linguistics research supposedly concentrates on domain-independent semantic representations, which is deemed "wrong," as it would not do justice to domain distinctions and would not enable the connection to domain-specific ontologies to allow for domain-specific reasoning. Regretting that a principled use of ontologies in NLP is not widespread, the authors finally paint a vision of an ontology-based NLP ecosystem modeled after their approach. While proposing the use of statistical methods to overcome brittleness, lack of coverage, and the limitations of logical

reasoning in the presence of uncertainty, it remains entirely unclear how this could be carried out.

Throughout, the book contains exercises that greatly help in internalizing the discussed concepts. Also, an array of available resources including a demo of the question answering system is announced. Some of these, however, were not yet available on the companion Web site at the time of review.

Even though the book states that its focus is "certainly not on robustness and coverage" (p. 6), I am still missing more concrete ideas on how robustness and coverage might be addressed than "by using machine learning techniques" (p. 142) without further elaboration. Thus, it is no coincidence that the implementation remains on the level of toy examples, however illustrating they might be. Also, while admitting to domain drift and the necessity of adaptation, it remains unclear how this adaptation, the integration of domain-specific and domain-independent processing, as well as the integration of several domains can be tractably attained. A deeper account on related approaches to computational semantics and a more recent bibliography on NLP/CL approaches could probably have avoided the impression that the interpretation of language as advocated here provides hardly more than other, same-old classic AI approaches.

Also, the pledge to the field of NLP to use more of Semantic Web technologies and services is more Semantic Web–centric than I would have expected from a book "attempting to provide a step towards the synergy between these two fields" (p. xv): True, it is unfortunate that these two fields interact as little as they do, as both are concerned with understanding the semantics of text, but the authors' clearly voiced credo on how NLP should finally follow suit is not very helpful for creating such synergy. To bridge the gap, the Semantic Web community should probably start leveraging results and evaluation methodologies from CL/NLP instead of reinventing/ignoring them, and the NLP community probably has to understand that ontological grounding and resolving entities to URIs is in fact enabling applications that go well beyond dependency parsing and word sense disambiguation. The Semantic Web community should understand that statistical methods and unsupervised acquisition are not the enemy, but the key to scalable, domain-adaptive processing of natural language, which has been known to evade total formalization since Sapier's "all grammars leak." The NLP community, especially statistical semantics, in turn, should understand the need and the demand for linking concepts to manually curated taxonomies and controlled vocabularies, as these are pervasive in industrial knowledge management. NLP should see the benefit in standardization for interoperability, and Semantic Webbers have to finally figure out that from format does not follow function.

To sum up, a well-written book goes the extra mile to exemplify its core contribution, which is to consequently drive language processing by the ontology of a target application from the very start. The exercises, together with the online materials, make it a useful resource for teaching. Unlike a similar proposal by Nirenburg and Raskin (2004), it is constructed and exemplified with Semantic Web technology such as RDF, OWL, SPARQL, open data, standardization, reasoning, and domain ontology. For this reason, the book will supposedly be very well received in the Semantic Web community, but could be perceived controversially by the NLP community: While the idea to radically align natural language processing to a target-domain ontology is reasonable, well-motivated from the application point of view, and a worthwhile contribution, the book unfortunately does not succeed in convincing most modern computational linguists that this approach can overcome known limitations of previous similar rule-based and knowledge-driven approaches, such as brittleness, poor scalability, and little adaptivity.

References

Bos J. 2008. Wide-coverage semantic analysis with boxer. *Proceedings of the 2008 Conference on Semantics in Text Processing STEP-08*, pages 277–286, Venice.

Cimiano P. 2009. Flexible semantic composition with DUDES. *Proceedings of the 8th International Conference on Computational Semantics (IWCS'09)*, pages 272–276, Tilburg.

Kamp H. and U. Reyle. 1993. *From Discourse to Logic. Introduction to Modeltheoretic Semantics of Natural Language, Formal Logic and Discourse Representation Theory*. Kluwer, Dordrecht.

Nirenburg S. and V. Raskin. 2004. *Ontological Semantics*. MIT Press, Cambridge, MA.

Schabes Y. 1990. *Mathematical and Computational Aspects of Lexicalized Grammars. Ph.D. thesis, University of Pennsylvania,* Philadelphia, PA.

This book review was edited by Pierre Isabelle.

Chris Biemann is Juniorprofessor (assistant professor) for Language Technology at Technische Universität Darmstadt, Germany. His research interests include statistical semantics, graph-based methods for unsupervised acquisition of semantic models, digital humanities, and cognitive computing. Chris's e-mail address is `biem@cs.tu-darmstadt.de`.