Ontologies and Lexical Resources for Natural Language Processing

To be published in

Cambridge Studies in Natural Language Processing

By

Cambridge University Press

Editors:

Chu-Ren Huang, Nicoletta Calzolari, Aldo Gangemi, Alessandro Lenci, Alessandro Oltramari, and Laurent Prevot

Expected year of publication: 2007

Content of Book:

Original Proposal

Motivation:

The interface between knowledge representation (KR) and natural language processing (NLP) is fast becoming a focal point of both fields. Ontologies have a special role to play in this interface. They are essential step stones (i) from natural language to knowledge understanding and manipulation and (ii) from formal theories of knowledge to their application in natural language processing. Moreover the emergence of the Semantic Web constitutes a unique opportunity to bring research result in this area to real world applications, at the leading edge of language engineering.

The most important area of this KR-NLP interface is between ontologies and lexical resources. On one hand, their integration includes, but is not restricted to, the use of ontologies (i) as language independent structures of multilingual computational lexicons, and (ii) as powerful tools for improving the performance of existing lexical resources on various NLP tasks such as word sense disambiguation. On the other hand, lexical resources constitute a formidable source of information for generating ontological knowledge both at foundational and domain levels.

The combination of ontology and lexical resources may also have implications for other fields. In essence, the combination of ontology and lexical resources gives us an inventory of human atomic concepts that can be further explored in cognitive studies and interface oriented applications.

Given the potential for this new research direction, there is a definite need for a general reference book on ontology and lexical resources. To the best of our knowledge, even though there are an increasing number of books on the study of ontology in related fields, there is no such book available in NLP. The recent publication of Nirenberg and Raskin's book on *Ontological Semantics (MIT Press)* gives an in-depth treatment of one approach, but lacks an overall perspective. On the other hand, the forthcoming book of *Ontolinguistics (Mouton)*, edited by Schalley and Zaefferer, focuses on the foundational issues of how ontologies affect linguistic concept encoding. Lastly, the number of available literature on ontologies is quickly increasing in neighboring fields, such as the book on ontologies and information system (Springer) by Kishore et al.

In sum, our book will serve dual purposes: The first is to provide general and indispensable reference to the study of ontology and NLP, while the second is to present an up-to-date state of the field, which will also serve as a reference point for future research.

References:

Kishore, Rajiv, Ram Ramesh, and Raj Sharman Eds. Forthcoming (2006). Ontologies in the Context of Information Systems. Berlin: Springer.

Sergei Nirenburg and Victor Raskin. 2004. Onotological Semantics. Cambridge: MIT Press.

Schalley, Adrea C., and Dietmar Zaefferer. Eds. Forthcoming (2006). *Ontolinguistics. How Ontological Status Shapes the Linguistic Coding of Concepts*. Berlin/New York: Mouton de Gruyter.

Targets of Publication

I. Audience:

Community: The NLP research community who are interested in ontology, as well as the Semantic Web and knowledge representation community who are interested in NLP. This will be a good fit for the target audience of CUP-SNLP. It will answer a clear and imminent need for literature on ontology for the NLP community.

Level: Post-graduate students and above

Abstracts of Chapters in the Book

I. Fundamental Aspects

I.1. The Lexical Representation of Knowledge: A multi-disciplinary perspective

Editors

This an original overview paper by the editors that establish the frame of reference for research on NLP using ontology and lexical resources. In essence, we adopt a mental lexicon approach where knowledge is lexically represented and accessed. Hence lexical resources represent the repertoire of human knowledge, while ontology imposes a structure for representation and reasoning. We will lay out the ground for the issues to be discussed in the book:

- -How knowledge is captured and represented by ontologies
  - How can lexical knowledge be converted to ontological knowledge
  - How can ontological representations be verified/enhanced by lexical knowledge,
  -How can ontological knowledge be applied to NLP

This overview paper will conclude with a road map for future research on the synergy of ontology and lexical resources.

I.2. SUMO and WordNet

Adam Pease and Christiane Fellbaum

I.3. DOLCE and Wordnet

Stefano Borgo, Aldo Gangemi, and Alessandro Oltramari

I.4. FrameNet and SUMO

Jan Scheffczyk, Collin F. Baker, Srini Narayanan

I.5 A roadmap for the ontologies lexical resources interface

Aldo Gangemi, Alessandro Oltramari

The OntoLex approach relies crucially on a comprehensive and well-structure lexical representation of knowledge. Hence, all the fundamental work in this field starts with synergizing a ontology and a lexical resources. The ontology chosen must be well-construed and robust, while the lexical resources must be comprehensive

intra-linguistically and robust enough to allow inter-lingual exchanges. We will introduce two of the most widely used ontology, with special emphasize on how they synergize with the most commonly used lexical resources: WordNet.

I.2.1. Introduction.

Nicoletta Calzolari

I.2.2. SUMO and WordNet

Adam Pease and Christiane Fellbaum

This paper introduces the first full-scale mapping between ontology and lexical resources – the mapping between SUMO and WordNet. Methodological and theoretical issues encountered in the mapping will be discussed. The main foci are on the implications of synergized ontology-lexicon for future studies in language technology and cognitive science.

I.2.3. DOLCE and Wordnet

Stefano Borgo, Aldo Gangemi, and Alessandro Oltramari

This paper discusses the general problems related to the semantic interpretation of WordNet taxonomy in the light of rigorous ontological principles inspired to the philosophical tradition, and introduce the *DOLCE* upper level ontology, which is inspired to such principles, yet with a clear orientation towards language and cognition.

II. Discovery and Representation of Conceptual Systems

II.1 Introduction

Aldo Gangemi

II.2. Experiments of Ontology Construction with Formal Concept Analysis

Sujian Li, Qin Lu, and Wenjie Li

This paper represents an ontology formally by type-subtype hierarchical relations. The main issue involved in applying Formal Concept Analysis to construct an ontology is the selection of data sources to acquire the necessary lexical knowledge. Two experiments are carried out for contrastive study of this issue, one adopts HowNet lexicon and the other uses a large-scale corpus.

II.3. Ontology, Lexicon, and Fact Repository as Leveraged To Interpret Events of Change

Marjorie McShane, Sergei Nirenburg, Stephen Beale

II.4. Hantology: An Ontology based on Conventionalized Conceptualization

Ya-Min Chou, and Chu-Ren Huang

This paper proposes the Chinese writing system as a linguistic ontology. Practically, a machine-readable ontology (written in OWL-DL) is created from Chinese characters and radicals. This ontology is also mapped to an existing upper level ontology (SUMO). The paper shows that it is possible to derive a robust conceptual system from Chinese radicals. Moreover, the framework developed allows for a deep study of language variation, a topic that has been so far neglected in the domain of linguistic ontology.

II.5. TBA

Aldo Gangemi and Carola Catenacci

III. Interfacing Ontologies and Lexical Resources

III.1. A methodology classification

Laurent Prevot, Stefano Borgo, and Alessandro Oltramari

This introduction compares several methodologies for interfacing ontologies and lexical resources, and analyses major projects in this area. Our results show that different methodologies lead to systems with very different characteristics.

III.2. MANAGELEX and the Semantic Web

Monica Gavrila, Cristina Vertan

This article describes a lexicon management tool which can be used for mapping lexical knowledge on ontology concepts, regardless of their original encoding formats. The tool's architecture and functionality are represented with focus on the connection between ontologies and lexicons.

III.3. Sinica BOW: A bilingual ontological wordnet

Chu-Ren Huang, Ru-Yng Chang, and Shiang-bin Li
This article introduces a bilingual approach towards merging ontological and lexical resources. English-Chinese cross-lingual mapping for WordNet and SUMO are carried out independently. These mappings are coupled with the mapping between SUMO and WordNet, as well as linked to various lexical resources. The result is a versatile infrastructure for exploring cross-lingual and mono-lingual lexical knowledge.

III.4. Restructuring WorNet's Top-Level: The OntoClean approach
Alessandro Oltramari, Aldo Gangemi, Nicola Guarino, and Claudio Masolo
This paper proposes an analysis and an upgrade of WordNet's top-level synset taxonomy of nouns. Some principles from OntoClean methodology are applied to the ontological analysis of WordNet. The OntoClean methodology characterizes concepts appearing in an ontology in terms of formal meta-properties.

III.5. Semantic Lexicons: Between Terminology and Ontology
Paul Buitelaar
The traditional model of semantic lexicons assigns senses to lexical items (i.e. words and/or more complex terminology), in which the set of senses is mostly open-ended. Ontologies provide formal class definitions for sets of objects, which can be seen as a 'sense' for those lexical items that express such objects. The model we describe here aims at merging these two disparate views into a unified approach to lexical semantics and ontology-based knowledge representation.

III.6. A Plug-in Approach to merge Global and Specialized Linguistic Ontologies
Bernardo Magnini and Manuela Speranza

IV. Learning and using ontological knowledge
IV.1. Introduction
Alessandro Lenci

IV.2. The Omega Ontology
Andrew Philpot, Eduard Hovy, Patrick Pantel
This paper presents the Omega ontology: the result of the synthesis between two major existing lexical resources (WordNet and Mikrokosmos)

subordinated to a new upper model. Omega is a shallow, lexically oriented term taxonomy. The paper explains how such a resource is useful in a wide variety of applications such as question answering and information integration.

IV.3. Automatic Acquisition of Lexico-semantic knowledge for QA

Lonneke van der Plas , Gosse Bouma

This paper demonstrates how automatically acquired lexico-semantic knowledge can be used to boost the performance of an open-domain question answering system for Dutch. We improve the precision of our QA system considerably. General WH-questions and definition questions benefit most: average precision goes from .42 to .57.

IV.4. Toward Medical Ontology Using Natural Language Processing

Eiji Aramaki, Takeshi Imai, Masayo Kashiwagi, Masayuki Kajino, Kengo Miyo, Kazuhiko Ohe

This paper presents a method to estimate term relations and term classification, which are the basic structure for the medical ontology. First, relations between medical terms are extracted from a medical text by using two methods as follows: (1) capturing the head of a definition sentence and (2) capturing typical phrases which indicate hypernym relations. Next, the terms are classified based on the co-occurrence verbs

IV.5. Automatic Thai Ontology Construction and Maintenance System

*Asanee Kawtrakul, Mukda Suktarachan, Aurawan Imsombu*t

This paper proposes three methodologies to build and maintain automatically an ontology for Thai, extracting information (or mining) from technical corpora, a dictionary and thesaurus. To build an ontology, terms are extract from corpora with the help of a shallow Parser. Syntactic-semantic constraints and Named Entities are used for identification of ontological relations. To extract relational terms from dictionaries, a Task-Oriented Parser is used. Finally, the Broader/Narrower relation of the Domain Specific thesaurus, AGROVOC, are mapped to an IS-A relation. Finally, a tool for experts to check the consistency and to extend the ontology is also developed.

IV.6. Rendering Semantic Ontologies: Automatic Extensions to

UMLS through Corpus Analysis

James Pustejovsky, Anna Rumshisky, Jose Castano

This paper discusses first the utility and deficiencies of existing ontology resources for a number of language processing applications. It then describes a technique (semantic re-rendering) for increasing the semantic type coverage of a specific ontology, the National Library of Medicine's UMLS. The technique involves the use of robust finite state methods in conjunction with large-scale corpus analytics of the domain corpus.

IV.7. The Duden Ontology: An Integrated Representation of Lexical and Ontological Information

Melina Alexa, Holger Schauer, Werner Scholze-Stubenrecht

This article reports a data model developed for the representation of lexical knowledge for the Duden Ontology. The model is the result of a cooperation between the publishing house Duden and the software company *intelligent views*. Our general aim is to create an asset pool in which all the information present in the Duden dictionaries is integrated in order to support reusability for different print and electronic products, provide solutions for language technology applications as well as support the efficient maintenance of the Duden dictionary data.

IV.8. Tying the Knot: Ground entities, Descriptions, and Information Objects for Construction-based for Information Extraction

Robert Porzel, Vanessa Micelli, Hidir Aras and Hans-Peter Zorn