Extracting World and Linguistic Knowledge from Wikipedia

Simone Paolo Ponzetto
Dept. of Computational Linguistics
University of Heidelberg
Heidelberg, Germany
http://www.cl.uni-heidelberg.de/~ponzetto

Michael Strube
EML Research gGmbH
Schloss-Wolfsbrunnenweg 33
Heidelberg, Germany
http://www.eml-research.de/~strube

Overview

Many research efforts have been devoted to develop robust statistical modeling techniques for many NLP tasks. Our field is now moving towards more complex tasks (e.g. RTE, QA), which require to complement these methods with a semantically rich representation based on world and linguistic knowledge (i.e. annotated linguistic data). In this tutorial we show several approaches to extract this knowledge from Wikipedia. This resource has attracted the attention of much work in the AI community, mainly because it provides semi-structured information and a large amount of manual annotations. The purpose of this tutorial is to introduce Wikipedia as a resource to the NLP community and to provide an introduction for NLP researchers both from a scientific and a practical (i.e. data acquisition and processing issues) perspective.

Outline

The tutorial is divided into three main parts:

1. **Extracting world knowledge from Wikipedia**. We review methods aiming at extracting fully structured world knowledge from the content of the online encyclopedia. We show how to take categories, hyperlinks and infoboxes as building blocks for a semantic network with unlabeled relations between the concepts. The task of taxonomy induction then boils down to labeling the relations between these concepts, e.g. with isa, part-of, instance-of, located-in, etc. relations.

2. **Leveraging linguistic knowledge from Wikipedia**. Wikipedia provides shallow markup annotations which can be interpreted as manual annotations of linguistic phenomena. These 'annotations' include word boundaries, word senses, named entities, translations of concepts in many languages. Furthermore, Wikipedia can be used as a multilingual comparable corpus.

3. **Future directions**. Knowledge derived from Wikipedia has the potential to become a resource as important for NLP as WordNet. Also the Wikipedia edit history provides a repository of linguistic knowledge which is to be exploited. Potential applications of the knowledge implicitly encoded in the edit history include spelling corrections, natural language generation, text summarization, etc.

Target audience

This tutorial is designed for students and researchers in Computer Science and Computational Linguistics. No prior knowledge of information extraction topics is assumed.

Proceedings of NAACL HLT 2009: Tutorials, pages 7–8,
Boulder, Colorado, June 2009.

Speakers' bios

Simone Paolo Ponzetto is an assistant professor at the Computational Linguistics Department of the University of Heidelberg, Germany. His main research interests lie in the area of information extraction, knowledge acquisition and engineering, lexical semantics, and their application to discourse-based phenomena.

Michael Strube is group leader of the NLP group at EML Research, a privately funded research institute in Heidelberg, Germany. The NLP group focuses on the areas of semantics, pragmatics and discourse and applications like summarization and information extraction.