Using a Natural Language Understanding System to Generate Semantic Web Content

Akshay Java, University of Maryland - Baltimore County, USA
Sergei Nimeburg, University of Maryland - Baltimore County, USA
Marjorie McShane, University of Maryland - Baltimore County, USA
Timothy Finin, University of Maryland - Baltimore County, USA
Jesse English, University of Maryland - Baltimore County, USA
Anupam Joshi, University of Maryland - Baltimore County, USA

ABSTRACT

We describe our research on automatically generating rich semantic annotations of text and making it available on the Semantic Web. In particular, we discuss the challenges involved in adapting the OntoSem natural language processing system for this purpose. OntoSem, an implementation of the theory of ontological semantics under continuous development for over 15 years, uses a specially constructed NLP-oriented ontology and an ontological-semantic lexicon to translate English text into a custom ontology-motivated knowledge representation language, the language of text meaning representations (TMRs). OntoSem concentrates on a variety of ambiguity resolution tasks as well as processing unexpected input and reference. To adapt OntoSem's representation to the Semantic Web, we developed a translation system, OntoSem2OWL, between the TMR language into the Semantic Web language OWL. We next used OntoSem and OntoSem2OWL to support SemNews, an experimental Web service that monitors RSS news sources, processes the summaries of the news stories, and publishes a structured representation of the meaning of the text in the news story.

Keywords: information extraction; natural language processing; OWL; RDF; Semantic Web

INTRODUCTION

A core goal of the development of the Semantic Web is to bring progressively more meaning to the information published on the Web. An accepted method of doing this is by annotating the text with a variety of kinds of metadata. Manual annotation is time-consuming and error-prone. Moreover, annotations must be made in a formal language whose use may require considerable training and expertise. Developing interactive
tools for annotation is a problematic undertaking because it is not known whether they will be in actual demand. A number of Semantic Web practitioners maintain that the desire to have their content available on the Semantic Web will compel people to spend the time and effort on manual annotation. However, even if such a desire materializes, people simply will not have enough time either to annotate each sentence in their texts or annotate a subset at a semantic level that is sufficiently deep to be used by advanced intelligent agents that are projected as users of the Semantic Web alongside people.

The alternative on the supply side is, then, automatic annotation. Within the current state of the art, automatically produced annotations are roughly at the level attainable by the latest information extraction techniques—a reasonably good level of capturing named entities with a somewhat less successful categorization of such entities (e.g., deciding whether Jordan is used as the first name of an individual or a reference to the Hashemite kingdom). Extracting more advanced types of semantic information, for example, types of events (to say nothing about determining semantic arguments, “case roles” in AI terminology), is not quite within the current information extraction capabilities, though work in this direction is ongoing. Indeed, semantic annotation is at the moment an active subfield of computational linguistics, where annotated corpora are intended for use by machine-learning approaches to building natural language processing capabilities.

On the demand side of the Semantic Web, a core capability is improving the precision of the Web search, which will be facilitated by detailed semantic annotations that are unambiguous and sufficiently detailed to support the search engine in making fine-grained distinctions in calculating scores of documents. Another core capability is to transcend the level of document retrieval and, instead, return as answers to user queries specially generated pragmatically and stylistically appropriate responses. To attain this capability, intelligent agents must rely on very detailed semantic annotations of texts. We believe that such annotations will be, for all intents and purposes, complete text meaning representations, not just sets of semantic or pragmatic markers (and certainly not templates filled with uninterpreted snippets of the input text that are generated by the current information extraction methods).

To attain such goals, Semantic Web agents must be equipped with sophisticated semantic analysis systems that process text found on the Web and publish their analyses on the Web as annotations in a form accessible to other agents, using standard Semantic Web languages such as RDF and OWL. The Semantic Web will, thus, be useful for both human readers and robotic intelligent agents. The agents will benefit from the existence of deep semantic annotations in their application-oriented information processing tasks and also will be able to derive such annotations from text. People will not directly access the annotation (metadata) level but will benefit from higher-quality and better formulated responses to their queries.

This article describes initial work on responding to the needs and leveraging the offerings of the Web by merging knowledge-oriented natural language processing with web technologies to produce both an automatic annotation-generating capability and an enhanced Web service oriented at human users. The ontological-semantic natural language processing system OntoSem (Nirenburg & Raskin, 2001) provided the basis for the automatic annotation effort. In order to test and evaluate the utility of OntoSem on the Semantic Web, we have developed SemNews (Java, Finin & Nirenburg, 2006; Java, Finin & Nirenburg, 2005), a prototype application that monitors RSS feeds of news stories, applies OntoSem to understand the text, and exports the computed facts back to the Web in OWL. A prerequisite for this system integration is a utility for translating knowledge formats between OntoSem’s knowledge representation language and ontologies and those of the Semantic Web.

Since our goal is to continuously improve the service, the quality of OntoSem results and system coverage must be continuously enhanced. The Web, in fact, contains a wealth
Related Content

A New Instance-Based Approach for Ontology Alignment
[www.igi-global.com/article/a-new-instance-based-approach-for-ontology-alignment/145229?camid=4v1a](www.igi-global.com/article/a-new-instance-based-approach-for-ontology-alignment/145229?camid=4v1a)

Improving the Information Security of Collaborative Web Portals via Fine-Grained Role-Based Access Control
[www.igi-global.com/chapter/improving-information-security-collaborative-web/39184?camid=4v1a](www.igi-global.com/chapter/improving-information-security-collaborative-web/39184?camid=4v1a)
Modeling and Querying XML-Based P2P Information Systems: A Semantics-Based Approach
[www.igi-global.com/chapter/modeling-querying-xmlbased-p2p-information/30391?camid=4v1a](www.igi-global.com/chapter/modeling-querying-xmlbased-p2p-information/30391?camid=4v1a)

Ontology Extraction Using Views for Semantic Web
[www.igi-global.com/chapter/ontology-extraction-using-views-semantic/31196?camid=4v1a](www.igi-global.com/chapter/ontology-extraction-using-views-semantic/31196?camid=4v1a)