SemNews: A Semantic News Framework

Akshay Java, Tim Finin and Sergei Nirenburg
University of Maryland, Baltimore County, Baltimore MD 21250
{aks1,finin,sergei}@umbc.edu
http://semnews.umbc.edu
http://ebiquity.umbc.edu/project/semnews

Introduction

SemNews is a semantic news service that monitors different RSS news feeds and provides structured representations of the meaning of news. As new content appears, SemNews extracts the summary from the RSS description and processes it using OntoSem, which is a sophisticated text understanding system.

The OntoSem environment is a rich and extensive tool for extracting and representing meaning in a language independent way. OntoSem performs a syntactic, semantic, and pragmatic analysis of the text, resulting in its text meaning representation or TMR. The TMRs are represented using a constructed world model or an ontology that consists of about 8000 Concepts. The ontology is also supported by an *Onomasticon* (Nirenburg & Raskin 2005) of about 400K terms, which is a lexicon of proper names. The learned instances from the text are stored in a *fact repository* which essentially forms the knowledge base of OntoSem. As the news items get processed by SemNews, the fact repository and the TMR are translated and published in Semantic Web representation language OWL.

Although significant number of documents already exist on the Semantic Web in representations such as RDF and OWL (Li Ding *et al.* 2004), a vast majority of content on the Web remains as natural language text. By integrating language understanding agents into the Semantic Web through the SemNews framework, we have been able to demonstrate the potential of large scale semantic annotation and automatic metadata generation.

OntoSem belongs to a general class of traditional, frame-based knowledge representation systems. Migrating such a system to newer web based representations like OWL, poses certain challenges. While doing a complete and faithful translation of knowledge from OntoSem's native representation into OWL is not feasible, we found the problems to be manageable in practice for a large subset of OntoSem features (Java, Finin, & Nirenburg 2005).

Unlike information extraction tools that provide features such as named entity detection and basic noun phrase markup, the SemNews application aims to exports facts and learned instances of ontological concepts. By publishing these facts on the Semantic Web we make available much more information to agents. In addition, the system also indexes the OWL versions of the TMR in Swoogle and stores it in an RDF triple store. These are used to support several services enabling people and agents to semantically browse, query and visualize the stories in the collection.

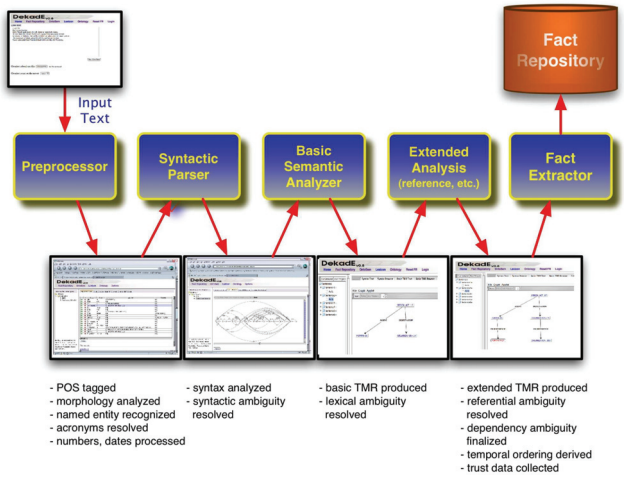

Figure 1: OntoSem goes through several basic stages in converting a sentence into a text meaning representation (TMR).

Architecture

As shown in Figure 1 the OntoSem environment takes as input unrestricted text and performs syntactic and semantic steps for extracting the meaning representation. The preprocessor deals with identifying sentence and word boundaries, part of speech tagging, recognition of named entities and dates, etc. The syntactic analysis phase identifies the various clause level dependencies and grammatical constructs of the sentence. The TMR is a representation of the meaning of the text and is expressed using the various concepts defined in the ontology.

The SemNews application, serves as a testbed for our work and has a simple architecture (Java, Finin, & Nirenburg 2006): (1) RSS from multiple sources is aggregated, parsed and then processed by the OntoSem (2) text processing en-

vironment. This results in the generation of TMRs (3) and updates to the fact repository (4). The Dekade environment (5) is a tool to edit the ontology and TMRs. OntoSem2OWL (6) converts the ontology and TMRs to their corresponding OWL versions (7,8). The TMRs are stored in the Redland triple (Beckett) store (9) and additional triples inferred by Jena (10).

There are also multiple viewers for searching and browsing the fact repository and triple store, enabling access to information that would otherwise not be easy to find using simple keyword based search. For example, one can browse through the story collection via the ontology to find stories that involve certain concepts, such as a 'terrorist organization' or 'assertive-act'. Since the ontology is linked to the lexicon, synonymous words such as 'tell' or 'say' would be resolved to the same ontological concept, i.e. ,assertive-act. OntoSem also provides named entity recognition and this interface allows users to find all stories that involve an entity in OntoSems onomasticon, such as *George Bush* or *Microsoft*. Each new named entity that the system finds is resolved for corefference within and across documents by looking up the fact repository.

The News Map feature allows visualization of the stories on a map based on the locations they reference. The text content of the news summary is indexed using lucene search engine and users may perform keyword searches.

SemNews supports RDQL or SPARQL interface to the triple store, which can be used in constructing arbitrary queries such as *find stories in which the nation named Afghanistan was the location of a bombing event*. Users can also define semantic alerts as queries over the RDF triple store and/or the Swoogle collection. For each alert, SemNews will generate an RSS feed of the results.

Preliminary Evaluation

OntoSem2OWL uses the Jena Semantic Web Framework (McBride 2001) internally to build the OWL version of the Ontology. The ontologies generated were successfully validated using two automated RDF validators: the W3C's RDF Validation Service [1] and the WonderWeb OWL Ontology Validator [2].

There were a total of about 8000 concepts in the original OntoSem ontology of which 7747 were successfully translated. The total number of triples generated was just over 100,000. These triples included a number of blank nodes – RDF nodes representing objects without identifiers that are required due to RDF's low-level triple representation.

Using the Jena API it takes about 10-40 seconds to build the model, depending upon the reasoner employed. The computation of transitive closure and basic RDF Schema inferencing takes approximately ten seconds on a typical workstation. The OWL Micro reasoner takes about 40 seconds while OWL Full reasoner fails, possibly due to the large search space.

[1] http://www.w3.org/RDF/Validator/

[2] http://phoebus.cs.man.ac.uk:9999/OWL/Validator

Related Work

Information extraction systems have been typically used to find named entities in text (Grishman & Sundheim 1996). Using OntoSem we go beyond entity recognition and extract the meaning of the text and relations in which these entities participate. The TAP (R.V.Guha & McCool 2003) project is an example of a system that uses simple information extraction technologies and represents the results in RDF. It uses a shallow but broad knowledge base containing basic lexical and taxonomic information about a wide range of popular objects. An example of another approach is a system developed by the Haystack Project (Hogue & Karger 2005). This semi-automated system enabled users to train a browser to extract Semantic Web content from HTML documents. The Cyc project has developed a very large knowledge base of common sense facts and reasoning capabilities. Recent efforts (Witbrock *et al.* 2004) include the development of tools for automatically annotating documents and exporting the knowledge in OWL.

Conclusion and Work in Progress

By mapping meaning extracted by automated language processing systems, we can provide agents on the Semantic Web with live and updated information. This also has an implication that agents on the Semantic Web should be able to reason in presence of incomplete or sometimes even inaccurate annotations. It is quite likely that for any KR system, transforming from one representation to another would always be lossful. However for most applications it would suffice even if partial transformations are provided.

The SemNews system is currently a research prototype that is being used to refine the underlying technologies and to explore how the sophisticated automatic linguistic processing of text can be integrated into the Semantic Web and conventional web applications. Ongoing work on SemNews includes an evaluation of its semantic recall and precision as well as a service that can group and cluster stories based on their semantic representations.

References

Beckett, D. Redland rdf application framework. `http://www.redland.opensource.ac.uk/docs`.

Grishman, R., and Sundheim, B. 1996. Message understanding conference-6: a brief history. In *Proceedings of the 16th Conference on Computational Linguistics*, 466–471.

Hogue, A., and Karger, D. R. 2005. Thresher: Automating the unwrapping of semantic content from the world wide web. In *Proceedings of the Fourteenth International World Wide Web Conference*.

Java, A.; Finin, T.; and Nirenburg, S. 2005. Integrating language understanding agents into the semantic web. In Payne, T., and Tamma, V., eds., *Proceedings of the AAAI Fall Symposium on Agents and the Semantic Web*.

Java, A.; Finin, T.; and Nirenburg, S. 2006. Text understanding agents and the Semantic Web. In *Proceedings of the 39th Hawaii International Conference on System Sciences*.

Li Ding, T. F.; Joshi, A.; Pan, R.; Cost, R. S.; Peng, Y.; Reddivari, P.; Doshi, V. C.; and Sachs, J. 2004. Swoogle: A search and metadata engine for the semantic web. In *Proceedings of the Thirteenth ACM Conference on Information and Knowledge Management*.

McBride, B. 2001. Jena: Implementing the RDF model and syntax specifi cation. In *Proceedings of the WWW2001 Semantic Web Workshop*.

Nirenburg, S., and Raskin, V. 2005. *Ontological semantics*. MIT Press.

R.V.Guha, and McCool, R. 2003. TAP: A semantic web toolkit. *Semantic Web Journal*.

Witbrock, M.; Panton, K.; Reed, S.; Schneider, D.; Aldag, B.; Reimers, M.; and Bertolo, S. 2004. Automated OWL Annotation Assisted by a Large Knowledge Base. In *Workshop Notes of the 2004 Workshop on Knowledge Markup and Semantic Annotation at the 3rd International Semantic Web Conference ISWC2004*.