OntoLT: A Protégé Plug-In for Ontology Extraction from Text

Paul Buitelaar, Daniel Olejnik, Michael Sintek

DFKI GmbH
Saarbruecken/Kaiserslautern, Germany
{paulb,olejnik,sintek}@dfki.de

1 Motivation

Ontologies are views of the world that tend to evolve rapidly over time and between different applications. Currently, ontologies are often developed in a specific context with a specific goal in mind. However, it is ineffective and costly to build ontologies for each new purpose each time from scratch, which may cause a major barrier for their large-scale use in knowledge markup for the Semantic Web. Creating ambitious Semantic Web applications based on ontological knowledge implies the development of new, highly adaptive and distributed ways of handling and using knowledge that enable existing ontologies to be adaptable to new environments.

As human language is a primary mode of knowledge transfer, a growing integration of language technology tools into ontology development environments is to be expected. Language technology tools will be essential in scaling up the Semantic Web by providing automatic support for ontology monitoring and adaptation. Language technology in combination with approaches in ontology engineering and machine learning provides linguistic analysis and text mining facilities for ontology mapping (between cultures and applications) and ontology learning (for adaptation over time and between applications).

2 Approach

The OntoLT approach provides a plug-in for the widely used Protégé ontology development tool, with which concepts (Protégé classes) and relations (Protégé slots) can be extracted automatically from annotated text collections. For this purpose, the plug-in defines a number of linguistic and/or semantic patterns over the XML-based annotation format that will automatically extract class and slot candidates. Alternatively, the user can define additional rules, either manually or by the integration of a machine learning process.

2.1 Linguistic/Semantic Annotation

The MM annotation format that is used by the OntoLT system integrates multiple levels of linguistic and semantic analysis in a multi-layered DTD, which organizes each level as a separate track with options of reference between them via indices [Vintar et al., 2002]. Linguistic/semantic annotation in the MM format covers: tokenization, part-of-speech tagging (noun, verb, etc.), morphological analysis (inflection, decomposition), shallow parsing (phrases, grammatical functions: subject, object, etc.) and lexical semantic tagging (synonyms) using EuroWordNet [Vossen, 1997].

2.2 Ontology Extraction From Text with OntoLT: An Example

Consider the development of an ontology for the computer science field from a corpus of relevant text documents (i.e., scientific papers). From this corpus we could, for instance, automatically extract and represent the occurring classes of technology (e.g., “web services”, “P2P platforms”, “RDF parsing”). In fact, this knowledge can be extracted from such sentences as: *…university develops P2P platform…*; *… University is the first group to develop an open source P2P platform…* By selecting the **`Institute-Verb-Obj`** pattern, the system selects all subjects of semantic class **`Institute`** (i.e., *university*) and extracts the corresponding verbs. By selecting one or more appropriate verbs (e.g., *develop, design, implement*), the user is presented with a list of automatically generated Protégé classes corresponding to the extracted objects of these verbs. Additionally, each of these classes will be assigned a slot **`institute`** of class **`Institute`**.

This extraction process is implemented as follows. OntoLT introduces a class called **`Mapping`** where the user can define the structure of the new classes and instances to be extracted. Each **`Mapping`** has **`Conditions`** and **`Operators.`** The **`Conditions`** describe the constraints that have to be fulfilled to be a candidate. The **`Operators`**

describe in which way the ontology should be enlarged if a candidate is found.

3 Related Work

A number of systems have been proposed for ontology extraction from text, e.g.: ASIUM [Faure et al., 1998], TextToOnto [Maedche and Staab, 2000], Ontolearn [Navigli et al., 2003]. Most of these systems depend on shallow text parsing and machine learning algorithms to find potentially interesting concepts and relations between them. The OntoLT approach is most similar to the ASIUM system, but relies even more on linguistic/semantic knowledge through its use of built-in patterns that map possibly complex linguistic (morphological analysis, grammatical functions) and semantic (lexical semantic classes, predicate-argument) structure directly to concepts and relations. A machine learning approach can easily be build on top of this but is not strictly necessary. Additionally, like the TextToOnto system, OntoLT provides a complete integration of ontology extraction from text into an ontology development environment, but selects for this purpose (unlike TextToOnto) the widely used Protégé tool, which allows for efficient handling and exchange of extracted ontologies (e.g., in RDF/S format).

Acknowledgements

This research has in part been supported by EC grants IST-2000-29243 for the OntoWeb project and IST-2000-25045 for the MEMPHIS project.

References

[Faure et al., 1998] Faure D., Nédellec C. and Rouveirol C. *Acquisition of Semantic Knowledge using Machine learning methods: The System ASIUM* Technical report number ICS-TR-88-16, 1998.

[Maedche and Staab, 2000] Maedche, A., Staab, S.: *Semi-automatic Engineering of Ontologies from Text.* In: Proceedings of the 12th International Conference on Software Engineering and Knowledge Engineering, 2000.

[Navigli et al., 2003] Navigli R., Velardi P., Gangemi A. *Ontology Learning and its application to automated terminology translation* IEEE Intelligent Systems, vol. 18:1, January/February 2003.

[Vintar et al., 2002] Vintar Š., Buitelaar P., Ripplinger B., Sacaleanu B., Raileanu D., Prescher D. *An Efficient and Flexible Format for Linguistic and Semantic Annotation* In: Proceedings of LREC, 2002.

[Vossen, 1997] Vossen P. *EuroWordNet: a multilingual database for information retrieval*. In: Proc. of the DELOS workshop on Cross-language Information Retrieval, March 5-7, Zürich, Switzerland.