Towards Ontology Engineering Based on Linguistic Analysis

Paul Buitelaar[1], Daniel Olejnik[1], Mihaela Hutanu[2], Alexander Schutz[2],

Thierry Declerck[2], Michael Sintek[3]

[1] DFKI GmbH, Language Technology
66123 Saarbruecken, Germany
{paulb,olejnik}@dfki.de

[2] Universität des Saarlandes, Computerlinguistik
66041 Saarbruecken, Germany
{hutanu, schutz, declerck}@coli.uni-sb.de

[3] DFKI GmbH, Knowledge Management
67608 Kaiserslautern, Germany
sintek@dfki.de

Abstract

In this paper we describe OntoLT, a plug-in for the widely used Protégé ontology development tool that supports the interactive extraction and/or extension of ontologies from text. The OntoLT approach aims at providing an environment for the integration of linguistic analysis in ontology development. OntoLT enables the definition of mapping rules with which concepts and attributes can be extracted automatically from linguistically annotated text collections. Mapping rules are defined by use of a constraint language. Constraints are implemented as XPATH expressions over the XML-based linguistic annotation. If all constraints are satisfied, the mapping rule activates one or more operators that describe in which way the ontology should be extended if a candidate is found.

Introduction and Related Work

Ontologies are formal, explicit specifications of shared conceptualizations, representing concepts and their relations that are relevant for a given domain of discourse (Gruber, 1994). With a recent increase in developments towards knowledge-based applications such as Intelligent Question-Answering, Semantic Web Services, Semantic-Level Multimedia Search, also the interest in large-scale ontologies has increased. To date, ontologies are mostly constructed completely by hand, which proves to be very ineffective and may cause a major barrier to their large-scale use in applications. Additionally, ontologies are domain descriptions that tend to evolve rapidly over time and between different applications (see e.g. Noy and Klein, 2002). Because of this, there has been an increasing development in recent years towards learning or adapting ontologies dynamically from related data (knowledge bases, databases, document collections).

Most of the work in ontology learning has been directed towards learning ontologies from text[1]. As human language is a primary mode of knowledge transfer, ontology learning from relevant text collections seems indeed a viable option as illustrated by a number of systems that are based on this principle, e.g. ASIUM (Faure et al., 1998), TextToOnto (Maedche and Staab, 2000) and Ontolearn (Navigli et al., 2003). All of these combine a certain level of linguistic analysis with machine learning algorithms to find potentially interesting concepts and relations between them.

A typical approach in ontology learning from text first involves the extraction of (more or less complex) terms from a domain-specific corpus. Extracted terms are statistically processed to determine their relevance for the domain corpus at hand and clustered into groups with the purpose of identifying a taxonomy of potential classes. Additionally, relations can be identified, mostly by computing a statistical measure of 'connectedness' between identified clusters.

Here we describe the OntoLT approach, which roughly follows a similar procedure. Additionally however, OntoLT aims at directly connecting ontology engineering with linguistic analysis through its use of mapping rules between linguistic structure and ontological knowledge (concepts and relations). In this way, linguistic knowledge (context words, morphological and syntactic structure, etc.) remains associated with the constructed ontology, which may be used subsequently in the application and maintenance of the ontology, e.g. in knowledge markup, ontology mapping and ontology evolution.

OntoLT

The OntoLT approach (introduced in Buitelaar et al., 2003) aims at providing an environment for the integration of linguistic analysis (of domain-specific corpora) in ontology engineering. OntoLT is available as a plug-in for the widely used Protégé ontology development tool[2] and enables the definition of mapping rules with which concepts (Protégé classes) and attributes (Protégé slots) can be extracted automatically from linguistically annotated text collections. A number of mapping rules are included with the plug-in, but alternatively the user can

[1] See for instance the overview of ontology learning systems and approaches in OntoWeb deliverable 1.5 (Gomez-Perez et al., 2003).

[2] http://protégé.stanford.edu

define additional rules, either manually or by the integration of a machine learning process.

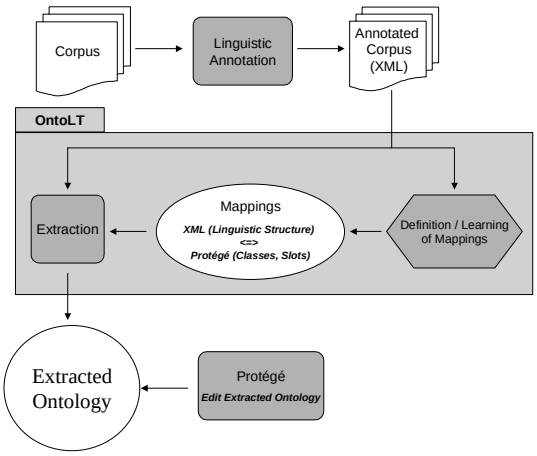

Figure 1: Overview of the OntoLT Approach

Mapping Rules

The ontology extraction process is implemented as follows. OntoLT provides a constraint language, with which the user can define constraint-based mapping rules. Constraints are implemented as XPATH expressions over the XML-based linguistic annotation. If all constraints are satisfied, the mapping rule activates one or more operators that describe in which way the ontology should be extended if a candidate is found.

Constraint Language

OntoLT provides a constraint language for defining mapping rules, which allows for the selection of particular linguistic entities in the annotated documents. Predefined constraints select for instance the predicate of a sentence, its linguistic subject or direct object. Constraints can also be used to check certain conditions on these linguistic entities, for instance if the subject in a sentence corresponds to a particular lemma. The constraint language consists of Terms and Functions, where Terms can be one of `AND, OR, EQUAL, NOT` and Functions can be one of `ID` (identify XML-node), `containsPath` (check if path exists) `hasLemma`, or `hasConcept` (check semantic class[3]).

Operators

Selected linguistic entities may be used in constructing or extending an ontology. For this purpose, OntoLT provides operators to create classes, slots or instances:

- `CreateCls:` create a new class
- `AddSlot:` add a slot to a class or create it if non-existing
- `CreateInstance:` introduce a new instance for an existing or new class
- `FillSlot:` set the value of a slot of an instance

OntoLT executes all mapping rules collectively. Therefore, according to which constraints are satisfied, all corresponding operators will be activated to create a set of candidate classes and slots that are to be validated by the user. According to this interactive process, classes and slots will be automatically generated into a new ontology or integrated into an existing ontology.

Linguistic Analysis

Linguistic annotation is not integrated with the Plug-In but is accessed via an XML-based exchange format, which integrates multiple levels of linguistic and semantic analysis in a multi-layered DTD with each analysis level (e.g. morphological, syntactic and dependency structure) organized as a separate track with options of reference between them via indices[4].

Linguistic annotation is currently provided by SCHUG, a rule-based system for German and English analysis (Declerck, 2002) that implements a cascade of increasingly complex linguistic fragment recognition processes. SCHUG provides annotation of part-of-speech (through integration of TnT: Brants, 2000), morphological inflection and decomposition (based on Mmorph: Petitpierre and Russell, 1995), phrase and dependency structure (head-complement, head-modifier and grammatical functions).

In Figure 2. below, we present a section of the linguistic annotation for the following sentence (German with corresponding sentence from the English abstract):

An 40 Kniegelenkpräparaten wurden mittlere Patellarsehnendrittel mit einer neuen Knochenverblockungstechnik in einem zweistufigen Bohrkanal bzw. mit konventioneller Interferenzschraubentechnik femoral fixiert.

(In 40 human cadaver knees, either a mid patellar ligament third with a trapezoid bone block on one side was fixed on the femoral side in a 2-diameter drill hole, or a conventional interference screw fixation was applied.)

The linguistic annotation for this sentence consists of PoS and lemmatization information in the <text> level, phrase structure (including head-modifier analysis) in the <phrases> level and grammatical function analysis in the <clauses> level (in this sentence there is only one clause, but more than one clause per sentence is possible).

For instance, the direct object (**DOBJ**) in this sentence (or rather in clause **cl1**) covers the phrase **p2**, which in turn corresponds to tokens **t5** to **t10** (*mittlere Patellarsehnendrittel mit einer neuen Knochenverblockungstechnik*). As token **t6** is a German compound word, a morphological analysis is included that corresponds to lemmas **t6.l1, t6.l2, t6.l3**.

[3] Semantic class information may be provided by a lexical semantic resource, such as WordNet (Miller, 1995) for English or EuroWordNet (Vossen, 1997) for various other languages, or by a domain-specific thesaurus or ontology, such as MeSH (Medical Subject Headings) for the biomedical domain: http://www.nlm.nih.gov/mesh/meshhome.html

[4] The format presented here is based on proposals and implementations described in (Buitelaar et al., 2003) and (Buitelaar and Declerck, 2003).

```
<sentence id="s3" stype="decl" corresp="">

<clauses>
 <clause id="cl1" from="p1" to="p5" pred="p5"
         type="pass">
  <arg id="a1" type="SUBJ" phrase="none" />
  <arg id="a2" type="IOBJ" phrase="p1"/>
  <arg id="a3" type="DOBJ" phrase="p2" />
  <arg id="a4" type="PP_ADJ" phrase="p3"/>
 </clause>
</clauses>

<phrases>
 …
 <phrase id="p2" from="t5" to="t10" type="NP">
  <mod from="t5" to="t5" />
  <head from="t6" to="t6" />
  <mod_post from="t7" to="t10" />
 </phrase>
 …
</phrases>

<text>
 <token id="t1" pos="APPR" str="An">
  <lemma id="t1.l1">an</lemma>
 </token>
 <token id="t2" pos="CARD" str="40" />
 <token id="t3" pos="NN"
        str="Kniegelenkpraeparaten">
  <lemma id="t3.l1">Kniegelenk</lemma>
  <lemma id="t3.l2">Praeparat</lemma>
 </token>
 <token id="t4" pos="VAFIN" str="wurden">
  <lemma id="t4.l1">werden</lemma>
 </token>
 <token id="t5" pos="ADJA" str="mittlere">
  <lemma id="t5.l1">mittler</lemma>
 </token>
 <token id="t6" pos="NN"
        str="Patellarsehnendrittel">
  <lemma id="t6.l1">patellar</lemma>
  <lemma id="t6.l2">Sehne</lemma>
  <lemma id="t6.l3">Drittel</lemma>
 </token>
 …
 <token id="t19" pos="ADJD" str="femoral" />
 <token id="t20" pos="VVPP" str="fixiert">
  <lemma id="t6.l1">fixieren</lemma>
 </token>
 <token id="t21" pos="PUNCT" str="." />
</text>
</sentence>
```

Figure 2: Linguistic Annotation Example

Ontology Extraction from Text with OntoLT

In order to test our approach in a realistic setting, we defined the following experiment. Given a corpus of medical texts in the neurology domain, we applied the OntoLT tool in combination with linguistic annotation as described above to extract a basic ontology for this domain.

The neurology corpus that we used in the experiment is a section of the MuchMore bilingual medical corpus (English-German) that includes around 9000 scientific abstracts in various domains[5] with around 1 million tokens for each language (see also: Buitelaar et al., 2004). The neurology section consists of 493 abstracts.

Statistical Preprocessing

In order to use only extracted linguistic information that is relevant for the domain, the approach includes a statistical preprocessing step. Here we base our approach on the use of the χ^2 function as described in (Agirre et al., 2001) for determining domain relevance. This function computes a relevance score by comparison of frequencies in the neurology corpus with that of frequencies in the rest of the MuchMore corpus. In this way, word use in the neurology domain is contrasted with that of medicine in general.

The χ^2 function gives a good indication of relevance, but experiments showed that also absolute frequency is an important indication of relevance. We therefore additionally multiply the χ^2 score by absolute frequency to obtain a combined measure of frequency and relevance.

In the following table, the 10 topmost relevant NP-heads, -modifiers and predicates (heads of clauses) are given for the neurology corpus (German with English translations):

NP-heads	Dysgenesie (dysgenesia)
	Denkstörung (thought disorder)
	Epilepsie (epilepsia)
	Psychiater (psychiatrist)
	Aura (aura)
	Tremor (tremor)
	Asystolie (asystole)
	Dopaminfreisetzung (dopamine release)
	Obdachlose (homeless)
	Aphasie (aphasia)
NP-modifiers	schizophren (schizophrenic)
	epileptisch (epileptic)
	transkraniel
	paranoid (paranoid)
	neuroleptisch (neuroleptic)
	neuropsychriatisch (neuro psychiatric)
	serotonerg
	impulsiv (impulsive)
	intraventrikulär (intra ventricular)
	neuropsychologisch (neuro psychological)
Predicates	zuerkennen (to adjudicate, award)
	staerken (to boost, encourage, strengthen)
	sparen (to conserve, save)
	betreten (to enter)
	hervorbringen (to create, produce)
	befuerworten (to support, endorse, advocate)
	gebrauchen (to employ, use)
	begreifen (to apprehend, understand)
	ueben (to exercise, practice)
	imitieren (to copy, imitate, mimic)

Table 1: 10 topmost relevant Heads, Modifiers and Predicates in the Neurology corpus

[5] The MuchMore corpus is publicly available from: http://muchmore.dfki.de/resources1.htm

OntoLT Mappings

The information provided by the statistical analysis is now used in defining mappings between the XML linguistic annotation and Protégé classes and slots. Here we discuss two such mappings: example 1. maps a head-noun to a class and in combination with its modifier(s) to one or more subclass(es); example 2. maps a linguistic subject to a class, its predicate to a corresponding slot for this class and the direct object to the "range" of this slot.

Example 1.: HeadNounToClass_ModToSubClass
This mapping generates classes for all head-nouns that were determined to be statistically relevant for the domain. For instance, classes are generated for the head-nouns *Dysgenesie (dysgenesia)* and *Epilepsie (epilepsia)*. Further, for each of these, sub-classes are generated that represent more specific concepts. For the two classes just mentioned, the following sub-classes are generated:

Dysgenesie	
Dysgenesie_kortikal	*(cortical)*
Epilepsie	
Epilepsie_myoklonisch	*(myoclonic)*
Epilepsie_idiopathisch	*(idiopathic)*
Epilepsie_fokal	*(focal)*

Example 2.: SubjToClass_PredToSlot_DobjToRange
This mapping generates for all statistically relevant predicates a class for the head-noun of their subject with a slot generated for the predicate and a slot range generated for the direct object. For instance, consider the sentence:

Transitorische ischaemische Attacken imitieren in seltenen Fällen einfache fokale motorische Anfälle.
("Transient ischemic attacks mimicking in some cases simple partial motor seizures.")

In this case, a class is generated for the head-noun of the subject (attacke - attack) and for the head-noun of the direct object (anfall - seizure). Further, a slot (imitieren - mimic) is generated for the class attacke with the class anfall as its range (class of possible fillers for this slot).

Conclusions and Future Work

OntoLT provides a middleware solution in ontology development that enables the ontology engineer to bootstrap the ontology from a relevant document collection. Currently, OntoLT is used in combination with SCHUG for linguistic annotation, but it may be combined with other linguistic analysis tools as well. Also, other XML formats can be supported, although XPATH expressions used by OntoLT will then need to be redefined.

Future work includes: 1. including linguistic analysis over a web service; 2. integrating an information extraction approach for ontology population (identifying class instances); 3. defining an evaluation platform to evaluate extracted ontologies in a quantitative way.

Acknowledgements

This research has in part been supported by EC grants IST-2000-29243 for the OntoWeb project, IST-2000-25045 for the MEMPHIS project and IST-2001-34373 for the ESPERONTO project.

References

Agirre E., Ansa O., Martinez D., Hovy E. Enriching WordNet concepts with topic signatures. In: Proceedings NAACL WordNet Workshop, 2001.

Brants, T. TnT - A Statistical Part-of-Speech Tagger. In: Proceedings of 6th ANLP Conference, Seattle, 2000.

Buitelaar P., Declerck Th., Sacaleanu B., Vintar Š., Raileanu D., Crispi C. A Multi-Layered, XML-Based Approach to the Integration of Linguistic and Semantic Annotations. In: Proceedings of EACL 2003 Workshop on Language Technology and the Semantic Web (NLPXML'03), Budapest, Hungary, April 2003.

Buitelaar P. and Declerck Th. Linguistic Annotation for the Semantic Web. In: Handschuh S., Staab S. (eds.) Annotation for the Semantic Web, IOS Press, 2003.

Buitelaar P., Olejnik D. and Sintek M. OntoLT: A Protégé Plug-In for Ontology Extraction from Text In: Proceedings of the Demo Session of the International Semantic Web Conference ISWC-2003, Sanibel Island, Florida, October 2003.

Buitelaar P., Steffen D., Volk M., Widdows D., Sacaleanu B., Vintar Š., Peters S. and Uszkoreit H. Evaluation Resources for Concept-based Cross-Lingual Information Retrieval in the Medical Domain. In: Proceedings of LREC2004.

Declerck Th. A set of tools for integrating linguistic and non-linguistic information. Proceedings of the SAAKM workshop at ECAI, Lyon, 2002.

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.

Gomez-Perez A., and Manzano-Macho D. A Survey of Ontology Learning Methods and Techniques. Deliverable 1.5, OntoWeb Project, 2003.

Gruber T. Towards principles for the design of ontologies used for knowledge sharing. Int. Journal of Human and Computer Studies 43(5/6), 1994, 907-928.

Maedche A. Ontology Learning for the Semantic Web. The Kluwer International Series in Engineering and Computer Science, Volume 665, 2003.

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.

Miller, G.A. WordNet: A Lexical Database for English. Communications of the ACM 11. 1995.

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

Petitpierre, D. and Russell, G. MMORPH - The Multext Morphology Program. Multext deliverable report for the task 2.3.1, ISSCO, University of Geneva. 1995.

Skut W. and Brants T. A Maximum Entropy partial parser for unrestricted text. In: Proceedings of the 6th ACL Workshop on Very Large Corpora (WVLC), Montreal. 1998.

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, 1997.