OLE — A New Ontology Learning Platform

Vít Nováček[*] and **Pavel Smrž**
Faculty of Informatics / Faculty of Information Technology
Masaryk University / Brno University of Technology
Botanická 68a / Božetěchova 2
602 00 Brno, Czech Republic / 612 66 Brno, Czech Republic
xnovacek@fi.muni.cz, smrz@fit.vutbr.cz

Abstract

This paper presents OLE — a new platform for bottom-up generation and merging of ontologies. In contrast to other ontology-learning systems that are currently available, OLE can be characterized by the modular architecture enabling integrating and comparing various methods of the automatic acquisition of semantic relations. We introduce the architecture of the tool and discuss the methodology of the employed synthetic bottom-up approach. OLITE — the central component responsible for the automatic acquisition of semantic relations from texts is described in detail. The presented preliminary results prove the efficiency of the implemented framework. We also provide a brief comparative overview of other relevant approaches and outline the future work on representation of uncertain knowledge for ontology merging.

1 Introduction

In the field of computer science, an ontology is understood as a formal and machine readable representation of a concept set, stratified in classes and including some relations among particular concepts and their classes. Ontologies are able to provide a comprehensive representation of information related to a particular subdomain of human knowledge. Such a representation can be utilized for an efficient semantic querying upon the subdomain objects, resource relevance measurement, interoperability of different systems and many other tasks. Ontologies also play the major role in the Semantic Web vision.

The basic approach to the ontology building is the manual definition of domain conceptualization. This task is usually performed by a group of domain experts. Various elaborated tools support the work; the most popular ones are Protégé, WebODE and OntoEdit. A comprehensive survey of such ontology engineering frameworks can be found in (Arpirez *et al.* 03).

The manual creation of ontologies presents a tedious work, is error-prone and the results are often too subjective. Moreover, it is infeasible to organize a group of experts for each possible domain. This led to the idea of automatic extraction of ontologies from available resources.

A method that can be used for ontology acquisition from texts was sketched by M. A. Hearst in (Hearst 92). It is based on the automatic pattern-based extraction of particular semantic relations. The hyponymy or *is-a* relation serves as a basis for the natural sub-concept/super-concept hierarchy of ontology classes. The notion of the automatic extraction of hyponymical constructs from textual data can be adopted for any other semantic relation, although the applied techniques may differ. Methods based on token co-occurrence can be employed to gather sets of concepts belonging to the same class. Various modifications of these two generic techniques are presented in (P. Pantel 04).

The OLE platform introduced in this paper takes advantage of the automatic acquisition methods. It enables creating the core taxonomy of an ontology subdomain in the bottom-up manner, from ontologies with a very simple structure to more complex ones, in a continual iterative process. It is also able to extend, refine and update ontologies with respect to new data.

A miniontology for each input resource is created first. It consists of concepts and classes gained from the given resource. The miniontologies are integrated into the current ontology on the fly. The process of ontology merging and alignment embodies the application of uncertainty representation methods. The emerging BayesOWL framework (Y. Peng 05) — a probabilistic extension of OWL — provides tools for this task.

OLE differs from other ontology-learning systems also in its accent on modularity and flexibility. Virtually any method of automated knowl-

[*] We would like to express out thanks to Masaryk University and Czech Academy of Sciences for supporting our work by the research grant T100300419.

edge acquisition can be employed as an independent part of the OLITE module. Section 3 gives details describing such an integration. Presently, the method of pattern-based extraction of semantic relations along with dynamic pattern learning is examined.

The rest of the paper is organized as follows. The next section presents a brief overview of the OLE architecture. Section 3 discusses one of the essential parts of OLE — the OLITE module. The efficiency of the platform is demonstrated by the results given in Section 4. Sections 5 and 6 compare OLE with other available systems and indicate the future directions of our research.

2 OLE Architecture

2.1 Design Considerations

The design of OLE has been influenced by the need for autonomy, efficiency and precision of the resulting platform. The following list summarizes the major requirements:

- The tool should support the user-friendly interactive way of ontology acquisition, but also the fully automatic process of knowledge mining that can run without any human assistance.

- The efficiency of ontology acquisition is crucial, for the system will process gigabytes of data.

- The precision is preferred over the recall. Even if the number of the extracted conceptual structures will be relatively low (compared to the number of relations a human can identify in the same resource), it will be balanced by the extensive quantity of resources available.

- The relations between concepts stored in the resulting ontology need not to be precise — the explicit uncertain knowledge representation is one of the essential parts of OLE. The loss of exactness is balanced by the increased fuzzy precision of the whole process.

2.2 System Components

The modular architecture of OLE is given in Figure 1.

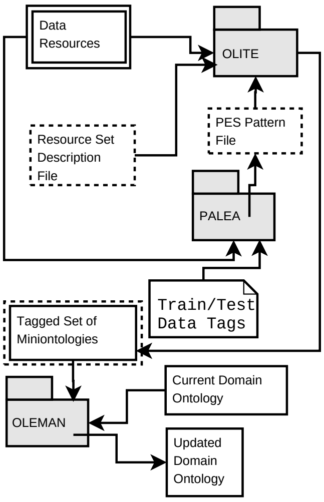

Figure 1: The modular architecture of the OLE platform

The **OLITE** module processes plain text and creates the miniontologies from the extracted data. The following sources are related to the OLITE module:

- *Data Resources* – documents provided by external tools (document classifiers, existing databases of related resources etc.).

- *Resource Set Description File* – a XML (RDF) encoded annotation of the resource files; it is read by the OLITE module in order to supply the extracted concept sets with category affiliation and other information.

- *PES Pattern File* – the definition of the semantic relation patterns.

- *Tagged Set of Miniontologies* – the output of the OLITE module in the form of miniontologies which correspond to the respective documents.

The **PALEA** module is responsible for the learning of new semantic relation patterns. A simple method of frequency analysis is integrated in the current implementation.

The **OLEMAN** merges the miniontologies resulting from the OLITE module and updates the base domain ontology. The uncertain information representation techniques are employed in the phase of ontology merging[1]. The module can

[1] Automatic matching of ontologies is a complex task which is not tackled in this paper. Relevant information on ontology merging and alignment can be found in (Doan *et al.* 03) and in (Ehrig & Staab 04). An introduction into the uncertain information in ontologies is given in (Y. Peng 05).

be used as a rudimentary ontology manager as well.

2.3 Implementation Remarks

All the OLE software components are implemented in the Python programming language. A special attention has been paid to the object oriented design. Another reason for choosing Python was the wide range of freely available relevant modules and application interfaces[2].

Python as an interpreted language can be inefficient for the implementation of some part of the OLE platform. Special tools improving the computational efficiency of the Python code are available. For example, we are going to take advantage of Psyco (Psy05) which is similar to the Java just-in-time compiler.

3 OLITE Module

3.1 Text Preprocessing

OLITE processes English plain-text documents and produces the respective miniontologies. To increase the efficiency, the input is preprocessed with the aim to reduce irrelevant data. Shallow syntactic structures that appear in the semantic pattern are also identified in this step.

The preprocessing consists of *splitting of the text into sentences*, *eliminating irrelevant sentences*, *text tokenization*, *POS tagging and lemmatization*, and *chunking*. The first two steps are based on regular expressions and performed in one pass through the input file. The possible relevance — the presence of a pattern — is detected by matching "core words" of the patterns.

The next phases of preprocessing depend on utilisation of NLTK natural language toolkit (NLT05) with custom-trained Brill POS tagging algorithm (Brill 94) and regular expression chunking incorporated. Moreover, the usage of NLTK toolkit (which allows users to train their own POS taggers from annotated data and easily create efficient chunking rules) enables to adapt the whole OLE system even for other languages than English in the future.

Fast regular expression-based chunking is then performed on the tagged sentences. The resulting file is stored in the form of a simple annotated vertical text — sentence elements with individual lines in the form of token/tag/lemma triples separated by the tab character. Tags conform to the Penn Treebank notation (see (Liu 05) for details).

3.2 Concept Extraction and Miniontology Generation

Any extraction algorithm can be integrated into OLITE in the form of a plug-in. Such a plug-in is responsible for the concept extraction, precise (or fuzzy) assignment of a class or a property and passing of gained information further in order to build an output miniontology.

The pattern-driven concept extraction process accepts patterns in a special form. The designed universal pattern-specification format allows new patterns to be easily added in the future.

The patterns are loaded and compiled from a separate PES file. PES stands for *Pattern Extended Specifications*. The syntax is similar to extended regular expressions, a few new symbols with higher level semantics are added. A chunk in a pattern is defined by a special expression — one of the NX, VX, AX, or UC character groups, representing a noun, verb, adjectival chunks or an unchunked text respectively. The expression can be amended by the '+' sign, indicating a sequence of same chunks. Chunk representation is enclosed in ˜ ˜. The core words are enclosed in % %. All other elements of the extended regular expressions syntax are accepted by the internal PES compiler.

The *is-a* pattern in the form of:

```
NP1 {‘‘,’’} ‘‘such as’’ NPList2
```

is transformed into the PES expression:

```
~NX~,? %such as% ~NX+~
```

Concept extraction utilizes the abstract regular expression matching again, but it works on the chunked sentences and the compiled PES patterns. The abstract matching means that the objects are not compared as standard strings. They carry information on what are they representing (a chunked sentence or a PES pattern) and what kind of operations should be applied.

The extracted information is stored in a universal internal format, no matter which extraction technique has been used. The output miniontology file is produced by applying respective translation rules. These rules are implemented as an independent plug-in (likewise the extraction algorithms) responsible for producing the output file

[2] For example, the NLTK natural language toolkit (NLT05) is used

in a desired format. Currently, the OWL DL format is supported only, but OLITE is able to produce any other format by the same mechanism.

4 Results

The method of pattern-based acquisition of simple relations was tested on general corpus texts containing about 10^8 words. The selected patterns are presented in the intuitive regular expression-like form in the first column of the table below[3].

The H_{abs} column contains numbers of matching sentences. Relative frequency of matches is given in the H_{rel} column[4]. The F_{all} field contains a ratio of successful pattern hits among randomly chosen sample of 50 matching sentences. Eventually, the F_{acq} column offers a ratio of conceptual structures acquirable by the OLITE module from the matching sentences.

Relatively high frequency of currently recognized semantic structures (compared to relations identified by a human) is very promising for further development. However, implementation of another techniques is essential in order to gain more general relations and even properties. Also, uniqueness of gained concepts must be examined properly across particular domains, because when choosing multidisciplinary random matches from the corpus, the measure is not very evidential.

5 Related Work

One of the best-known ontology-acquisition efforts is represented by the OntoLearn project (Gangemi *et al.* 03). The statistical methods based on frequency measures are equipped in terminology extraction from the source data in OntoLearn as well as in OLE. However, the systems considerably differ in their use of the "template" ontology. The WordNet database is queried in several stages of the semantic interpretation and specific relation discovery in OntoLearn. New relation patterns are inferred based on the known WordNet conceptual relations. Therefore, the results of OntoLearn are determined by the coverage of WordNet. On the other hand, the process of ontology acquisition can start from scratch in OLE and the current "template" ontology can be dynamically extended and refined.

The KnowItAll (Etzioni *et al.* 04) system incorporates the same extraction of semantic relations as is implemented in OLE. The uncertainty is introduced in the form of so called web-scale probability assessment in KnowItAll, but not as a part of the ontology structure itself. OLE represents the whole conceptual structure of a given domain in the unified system integrating the uncertain information.

The OLITE and PALEA modules implement just basic methods of pattern learning and their application. Advanced algorithms for pattern-based extraction of semantic relations are described in (Etzioni *et al.* 04), (Hearst 92) and (P. Pantel 04). The concept clustering techniques for terascale knowledge acquisition are introduced in (P. Pantel 04), (T. T. Quan 04) presents the fuzzy concept clustering. All these techniques can be adopted by the OLITE module to supplement the dynamic pattern learning and application.

6 Conclusions and Future Directions

OLE is primarily intended for autonomous creation and management of domain specific ontologies. The bottom-up approach to the ontology acquisition is emphasized, as well as the need for uncertainty representation. The preliminary results clearly show that OLE provides the modular and flexible platform for comparing and testing various information extraction techniques. The OLITE component implements the basic knowledge acquisition methods, other modules can be easily added.

Challenging work remains to be done on the PALEA module, especially in the area of dynamic acquisition of new patterns for additional semantic relations. Many advanced techniques for concept mining still wait for their implementation. One of them is FFCA — the Fuzzy Formal Concept Analysis (T. T. Quan 04) which is based on fuzzy concept clustering. The notion of uncertainty is implicitly embraced already at the initial level of information extraction. The ontology merging process in OLE will benefit from this approach. It will be implemented as another information extraction plug-in.

[3] The patterns are partially adopted from (Etzioni *et al.* 04) and (Hearst 92).

[4] The overlap among the matches was found to be insignificant.

Selected *isa* patterns	H_{abs}	H_{rel}	F_{all}	F_{acq}
NP (and|or) other NP	17384	0.28	94	85
NP including (NPList (and|or))? NP	23985	0.38	92	73
NP (is|was) a NP	140632	2.26	66	30
(NPList)? NP like NP	147872	2.37	16	14
sums (H fields) and averages (F fields)	329873	5.29	52.00	50.50

Table 1: The most productive extraction patterns

7 Acknowledgements

This work is partially supported by the Academy of Sciences of the Czech Republic, 'Information Society' program, the research grant T100300419.

References

(Arpirez *et al.* 03) J. C. Arpirez, O. Corcho, M. Fernandez-Lopez, and A. Gomez-Perez. Webode in a nutshell. *AI Mag.*, 24(3):37–47, 2003.

(Brill 94) E. Brill. A report of recent progress in transformation-based error-driven learning. In *Proc. ARPA Human Language Technology Workshop '94*, Princeton, NJ, 1994.

(Doan *et al.* 03) A. Doan, J. Madhavan, R. Dhamankar, P. Domingos, and A. Halevy. Learning to match ontologies on the semantic web. *The VLDB Journal*, 12(4):303–319, 2003.

(Ehrig & Staab 04) M. Ehrig and S. Staab. Qom - quick ontology mapping. In *ISWC 2004: Third International Semantic Web Conference. Proceedings*, pages 683–697, 2004.

(Etzioni *et al.* 04) O. Etzioni, M. Cafarella, D. Downey, S. Kok, A.-M. Popescu, T. Shaked, S. Soderland, D. S. Weld, and A. Yates. Web-scale information extraction in knowitall: (preliminary results). In *WWW '04: Proceedings of the 13th international conference on World Wide Web*, pages 100–110, New York, NY, USA, 2004. ACM Press.

(Gangemi *et al.* 03) A. Gangemi, R. Navigli, and P. Velardi. Corpus driven ontology learning: a method and its application to automated terminology translation. *IEEE Intelligent Systems*, pages 22–31, 2003.

(Hearst 92) M. A. Hearst. Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 14th conference on Computational linguistics*, pages 539–545, Morristown, NJ, USA, 1992. Association for Computational Linguistics.

(Liu 05) H. Liu. Montylingua: An end-to-end natural language processor with common sense. Available at: http://web.media.mit.edu/~hugo/montylingua, 2005.

(NLT05) *NLTK: Natural Language Toolkit – Technical Reports*, 2005. Available at: http://nltk.sourceforge.net/tech/index.html.

(P. Pantel 04) E. Hovy P. Pantel, D. Ravichandran. Towards terascale knowledge acquisition. In *Proceedings of Conference on Computational Linguistics (COLING-04)*, pages 771–777, 2004.

(Psy05) *The Ultimate Psyco Guide*, 2005. Available at: http://psyco.sourceforge.net/psycoguide/index.html.

(T. T. Quan 04) T. H. Cao T. T. Quan, S. C. Hui. Automatic generation of ontology for scholarly semantic web. In *ISWC 2004: Third International Semantic Web Conference. Proceedings*, pages 726–740. Springer-Verlag Berlin Heidelberg, 2004.

(Y. Peng 05) R. Pan Y. Peng, Z. Ding. Bayesowl: A probabilistic framework for uncertainty in semantic web. In *Proceedings of Nineteenth International Joint Conference on Artificial Intelligence (IJCAI05)*, 2005.