BOLE — A New Bio-Ontology Learning Platform

Vít Nováček
Faculty of Informatics
Masaryk University
Botanická 68a, 602 00 Brno, Czech Republic
xnovacek@fi.muni.cz

Pavel Smrž
Faculty of Information Technology
Brno University of Technology
Božetěchova 2, 612 66 Brno, Czech Republic
smrz@fit.vutbr.cz

ABSTRACT

This paper presents BOLE — a new platform for bottom-up generation and merging of bio-ontologies. In contrast to other ontology-learning systems that are currently available, BOLE 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 bio-ontology merging.

Keywords

ontology extraction, text mining

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. Bio-ontologies are able to provide a comprehensive representation of information related to a particular subdomain of biology, medicine, etc. 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 bio-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 [2].

The manual creation of bio-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 [9]. 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 [12].

The BOLE platform introduced in this paper takes advantage of the automatic acquisition methods. It enables creating the core taxonomy of a bio-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 bio-ontologies with respect to new data.

A miniontology for each input biomedical document is created first. It consists of concepts and classes gained from the given resource. The miniontologies are integrated into the current bio-ontology on the fly. The process of ontology merging and alignment embodies the application of uncertainty representation methods. The emerging BayesOWL framework [13] — a probabilistic extension of OWL — provides tools for this task.

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

Section 3 discusses one of the essential parts of BOLE — the OLITE module. The efficiency of the platform is demonstrated by the results given in Section 4. Sections 5 and 6 compare BOLE with other available systems and indicate the future directions of our research.

2. BOLE ARCHITECTURE

2.1 Design Considerations

The design of BOLE 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 biomedical 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 bio-ontology need not to be precise — the explicit uncertain knowledge representation is one of the essential parts of BOLE. The loss of exactness is balanced by the increased fuzzy precision of the whole process.

2.2 System Components

The modular architecture of BOLE is given in Figure 1.

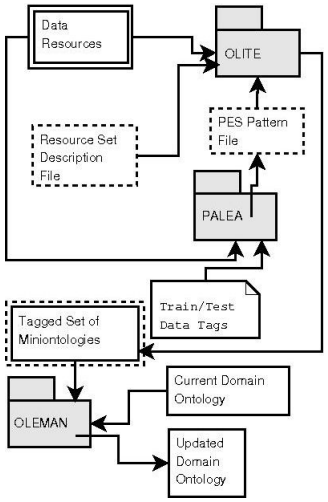

Figure 1: The modular architecture of the BOLE platform

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, e. g. MEDLINE documents.
- *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 bio-ontology. The uncertain information representation techniques are employed in the phase of ontology merging[1]. The module can be used as a rudimentary bio-ontology manager as well.

2.3 Implementation Remarks

All the BOLE 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 BOLE platform. Special tools improving the computational efficiency of the Python code are available. For example, we are going to take advantage of Psyco [14] 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.

[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 [5] and in [6]. An introduction into the uncertain information in ontologies is given in [13].

[2]For example, the MontyLingua natural language processing tool [11] is used

The next phases of preprocessing depend on a slightly modified Python module of the MontyLingua natural language processing toolkit [11]. MontyLingua incorporates the enriched Brill94 algorithm [3] for POS tagging and successive tasks.

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 [11] for details).

3.2 Biomedical 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 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 biomedical 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 [8]. The statistical methods based on frequency measures are equipped in terminology extraction from the source data in OntoLearn as well as in BOLE. 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 BOLE and the current "template" ontology can be dynamically extended and refined.

The KnowItAll [7] system incorporates the same extraction of semantic relations as is implemented in BOLE. 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. BOLE 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 [7], [9] and [12]. The concept clustering techniques for terascale knowledge acquisition are introduced in [12], [15] presents the fuzzy concept clustering. All

[3]The patterns are partially adopted from [7] and [9].

[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 patterns extracted from biomedical texts

these techniques can be adopted by the OLITE module to supplement the dynamic pattern learning and application.

6. CONCLUSIONS AND FUTURE DIRECTIONS

BOLE is primarily intended for autonomous creation and management of domain specific bio-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 BOLE 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 [15] 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 BOLE will benefit from this approach. It will be implemented as another information extraction plug-in.

7. ACKNOWLEDGEMENTS

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

8. REFERENCES

[1] J. Allen and G. Ferguson, Actions and Events in Interval Temporal Logic, In O. Stock (Ed.), Spatial and Temporal Reasoning, Kluwer Academic Publishers, pp. 205–245, 1997.

[2] J. C. Arpirez, O. Corcho, M. Fernandez-Lopez, A. Gomez-Perez. WebODE in a Nutshell. Available at `http://www.findarticles.com/p/articles/mi_m2483/is_3_24/ai_110575582`, 2005.

[3] A Report of Recent Progress in Transformation-Based Error-Driven Learning. Available at `http://www.ifi.unizh.ch/CL/volk/LexMorphVorl/Brill94.pdf`, 2005.

[4] Burnard, L., ed.: Users Reference Guide for the British National Corpus. Oxford University Computing Service (1995)

[5] A. Doan et al. Learning to Match Ontologies on the Semantic Web. In *The VLDB Journal (2003) 12*, pp. 303-319.

[6] M. Ehrig, S. Staab. QOM – Quick Ontology Mapping. In *Proceedings of The Semantic Web – ISWC 2004*, pp. 683-697. Springer-Verlag, Berlin Heidelberg 2004.

[7] O. Etzioni, et al. Web-Scale Information Extraction in KnowItAll (Preliminary Results). In *Proceedings of the 13th international conference on World Wide Web*, pp. 100-110. New York, NY, USA 2004.

[8] A. Gangemi, R. Navigli, P. Velardi. Corpus Driven Ontology Learning: a Method and its Application to Automated Terminology Translation. In *IEEE Intelligent Systems, January-February 2003*, pp. 22-31.

[9] M. A. Hearst. Automatic Acquisition of Hyponyms from Large Text Corpora. In *Proceedings of the Fourteenth International Conference of Computational Linguistics*. Nantes France, July 1992.

[10] J. Hobbs. A DAML Ontology of Time, `http://www.cs.rochester.edu/~ferguson/ daml/`, 2002.

[11] H. Liu. MontyLingua: An end-to-end natural language processor with common sense. Available at: `http://web.media.mit.edu/~hugo/montylingua`, 2005.

[12] P. Pantel, D. Ravichandran, E. Hovy. Towards Terascale Knowledge Acquisition. In *Proceedings of Conference on Computational Linguistics (COLING-04)*, pp. 771-777. Geneva, Switzerland 2004.

[13] Y. Peng, Z. Ding, R. Pan. BayesOWL: A Probabilistic Framework for Uncertainty in Semantic Web. Submitted to *Nineteenth International Joint Conference on Artificial Intelligence (IJCAI05)*. Available at: `http://userpages.umbc.edu/~zding1/pub/ijcai05Final.pdf`, 2005.

[14] Psyco Homepage. Available at `http://psyco.sourceforge.net`, 2005.

[15] T. T. Quan, S. C. Hui, T. H. Cao. Automatic Generation of Ontology for Scholarly Semantic Web. In *Proceedings of The Semantic Web – ISWC 2004*, pp. 726-740. Springer-Verlag, Berlin Heidelberg 2004.

[16] Y. Sure et al. OntoEdit: Collaborative Ontology Development for the Semantic Web. In *Proceedings of The Semantic Web – ISWC 2002*, pp. 221-235. Springer-Verlag, Berlin Heidelberg 2002.