

Language Resources for Semantic Document Annotation and Crosslingual Retrieval

Petya Osenova, Kiril Simov
Bulgarian Academy of Sciences
{petya|kivs}@bultreebank.org

Eelco Mossel
University of Hamburg
mossel@informatik.uni-hamburg.de

Abstract

1 Introduction

Given the huge amount of static and dynamic contents created for eLearning tasks, the major challenge for their wide use is to improve their accessibility within Learning Management Systems (LMS). The LT4eL project¹ tackles this problem by integrating semantic knowledge to enhance the management, distribution and retrieval of the learning material [1].

The semantic annotation has already become a key ingredient of Semantic Web. There is already a vast quantity of literature and initiatives, which approach this topic from various perspectives. For example, there was SAAW 2006 - the First Semantic Authoring and Annotation Workshop devoted on tools, standards and practice of semantic annotation. In this paper, we present a model of the relation of a domain ontology to text. In order to facilitate this relation we need to construct corresponding language resources including lexicons and grammars. Here we discuss the nature of these resources. Also we describe how they can be used for mono- and multilingual search.

The abstract is structured in the following way: first we present our model for relation of a domain ontology and domain text; then we present the ontology search based on this relation; the fourth section reports on a comparison between text-based search and ontology search; the last section concludes the paper.

2 Domain Ontology and Semantic Annotation

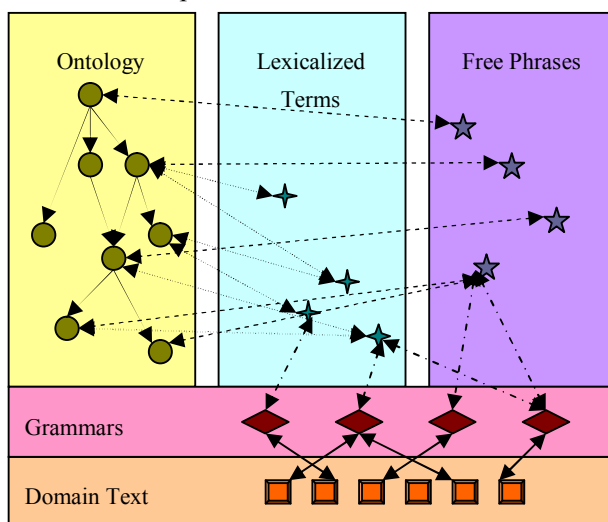
The domain of the learning corpus in the LT4eL Project is “Computer Science for Non-Computer Scientists”. It covers topics like operating systems; programs; document preparation – creation, formatting, saving, printing; Web, Internet, computer networks; HTML, websites, HTML documents; email, etc. The main application of the ontology is: the indexing of these domain documents with concept information and translation of the same information across different languages.

The creation of the ontology was done on the basis of manually annotated keywords in the eight languages of the project (Bulgarian, Czech, Dutch, English, German, Polish, Portuguese, Romanian). The annotated keywords from the other languages were translated into English. Then by search on the Web we collected definitions for the keywords. The set of definitions of a keyword highlights the various meanings of the keyword and the relations between its meanings and other concepts. After the determination of the keywords meanings we created concepts corresponding to these meanings. These concepts constitute the backbone of the domain ontology. The next step of the ontology development was to map the domain concepts to an upper ontology (in our case we used DOLCE – [2], [3]) in order to inherit some knowledge already encoded in the upper ontology (relations, for instance) and to ensure right concept classification with respect to concept metaproperties defined in the ontology creation methodology – OntoClean [4]. The mapping was facilitated by OntoWordNet [5]. Additionally the ontology was extended with concepts from other sources like terminological lexicons and Wikipedia. At the moment, the domain ontology contains about 900 domain concepts, about 50 concepts from DOLCE and about 200 intermediate concepts from OntoWordNet.

In order to use the ontology for semantic search over documents, we need to establish a connection between the ontology and the texts of the documents. We establish this connection by three language resources and tools. The

¹ <http://www.lt4el.eu/> – the LT4eL (Language Technology for eLearning) project is supported by the European Community under the Information Society and Media Directorate, Learning and Cultural Heritage Unit.

first one is a lexicon aligned to the ontology. The lexicon for each language contains lexical items grouped by their meaning which is represented in the ontology. There exist various attempts to approach this mapping task. Most of them start from lexicons already existing for several languages, and then try to establish a connection between the concepts defined in these lexicons. Such initiatives were WordNet [6], EuroWordNet [7], SIMPLE [8]. In spite of the fact that we employ the experience from these projects (mapping to WordNet and Pustejovsky's ideas in SIMPLE), we also suggest an alternative in connecting the ontology and the lexicons. Our model is very close to the LingInfo model (see [9]) with respect to the mapping of the lexical items to concepts, but also with respect to the other language processing tools we connect to the ontology – the concept annotation grammars and concept disambiguation tools. Thus, the other two language resources are (1) partial grammars which facilitate the mapping from the lexical items to their realization in the texts; and (2) disambiguation rules which solve the problem of ambiguity of the lexical items on the basis of the context of their usage in the texts. The model is graphically depicted in the next picture:



These mappings ensure the relation between ontology elements and their realization in the text. This relation is a basis for (1) monolingual search in which the ontology inference is used for query expansion; and for (2) multi-language search in which, in addition to query expansion, the ontology is used as a mediator between the different languages.

3 Ontology-Based Search

One of the goals of the project is to develop a search functionality which (1) improves the accessibility to documents in a learning management system by exploiting semantic characteristics of search queries and documents; (2) works for several languages and (3) enables users to find documents in several languages while using search terms or an ontology representation only in the user's language.

The search functionality is based on the following three resources as described above: the term-concept lexicons, the ontology and the concept annotation of the documents. Presuppositions for using the search are the availability of documents and a lexicon in at least one of the languages the user knows, and that the topics of the documents are (at least partly) covered by the ontology.

The basic idea of the ontology-based search is that concepts from the ontology lead the user to those documents which are appropriate for his query. The search will be most precise when the user directly selects concepts from the ontology. However, we start with a free-text query for two reasons. First, we assume that the users, who are probably familiar with Google, want their results fast, with not too many intermediate steps. The simplest case is to type search words and click on search - this procedure is also used for full-text search in our system. Second, we use the entered search words to find a good starting point in the ontology, so that the user does not have to click his way through the ontology starting at the root.

The entered words are looked up in the lexicon, and the concepts that are linked to the matching lexicon entries are used for ontology-based search in an automatic fashion. Before lexicon lookup, the words are orthographically normalised, and combinations for multi-word terms are created (e.g. if the words "text" and "editor" are entered, the combinations "texteditor", "text editor" and "text-editor" are created and looked up, in addition to the individual words). For each of the found concepts, the set of all its (direct or indirect) subconcepts is determined, and is used to retrieve documents.

When the found documents are displayed, at the same time the relevant parts of the ontology are presented. Now, in a second step, the user can select (by marking a checkbox) the concept(s) he wants to look for and repeat the search. If an entered word was ambiguous, the intended meaning can be explicated now by selecting the appropriate concept. Furthermore, by clicking on a concept, related concepts are displayed; navigation through the ontology is possible in this way, following the ontological relations.

A list of retrieval languages (only documents written in one of those languages will be found) is specified as an input parameter. The retrieved documents are sorted by language. The next ordering criterion is a ranking, based on the number of different search concepts and the number of occurrences of those concepts in the document. For each found document, its title, language, and matching concepts are shown.

4 Evaluation

In a small experiment, which is part of the evaluation of the search function, simple text search and semantic search have been compared. The basic task which for this evaluation was as follows: two terms (which were also lexical entries in the lexicon of the language under investigation) have been chosen as parts of a query (the equivalents of the terms "program" and "slides" in each of the languages²). We encode the query as a full text search (including inflected forms and word order variations), and as a semantic search using the ontology to expand the query with all subconcepts and one superconcept. These queries have been run on the document set. This resulted in two sets of paragraphs, one for the text search and one for the semantic search. The conceptual annotation has been used to identify the paragraphs in these documents. The conjunction of these two result sets has been investigated by a researcher and each paragraph rated as either relevant or irrelevant to the search. The retrieval results of both methods (text search and semantic search) have been weighted against the set of relevant paragraphs with the well-know measures of recall and precision. Both values have been combined in an F-measure. The F-measure is used to compare the results.

The experiment has been run for six languages: Bulgarian, Dutch, English, German, Polish and Portuguese. The F-measures for both text search and semantic search are presented in Table 1. The gain is due to improvements in both recall and precision. It is significant for all languages. The gain is the lowest for Portuguese, because there were only a small number of returned documents. Also, there is visible variation between the languages.

Language	Text Search	Semantic search
Bulgarian	56,25	91,30
Dutch	47,50	94,12
English	27,96	79,42
German	36,00	59,26
Polish	12,50	50,00
Portuguese	28,67	33,33

Table 1: F-measures for full text search and semantic search in six languages.

² The interpretation of the query is "Which program do you use for the preparation of slides?"

Another factor that played a role in the results was the context: the narrower the context (e.g. sentences), the better the results, and vice versa. As has been said before, the conceptual search produces results only in those cases where the search words are in the lexicon and thus matched to concepts in the ontology. This has been the case in the evaluation example. In the case where the search word does not match a lexical item, the text search as well as the keyword-based search is used as a fallback strategy.

It is subject to further user-centred evaluations to reveal how well received the semantic search is in the context of the learning management systems where it is used to solve real tasks. This part of the evaluation is on-going.

5 Conclusion

In this abstract, we described an approach for relating an ontology to a multilingual collection of texts, the employed resources and their usage for ontology-based search. We also present first steps of the evaluation of the search functionality, which show that the ontology-based search significantly improves the retrieval for several languages.

References

- [1] Paola Monachesi, Lothar Lemnitzer, Kiril Simov. Language Technology for eLearning. Proceedings of EC-TEL 2006, in Innovative Approaches for Learning and Knowledge Sharing, LNCS 0302-9743, pp. 667-672. 2006
- [2] Claudio Masolo, Stefano Borgo, Aldo Gangemi, Nicola Guarino, Alessandro Oltramari and Luc Schneider. 2002. The WonderWeb Library of Foundational Ontologies. WonderWeb Deliverable D17, August 2002. <http://www.loa-cnr.it/Publications.html>.
- [3] Claudio Masolo, Stefano Borgo, Aldo Gangemi, Nicola Guarino and Alessandro Oltramari. 2002. Ontology Library (final). WonderWeb Deliverable D18, December 2003. <http://www.loa-cnr.it/Publications.html>.
- [4] Nicola Guarino and Christopher Welty. 2002. "Evaluating Ontological Decisions with OntoClean." Communications of the ACM, 45(2): 61-65.
- [5] Aldo Gangemi, Roberto Navigli, Paola Velardi. The OntoWordNet Project: extension and axiomatization of conceptual relations in WordNet. Meersman R, et al. (eds.), Proceedings of ODBASE03 Conference, Springer, 2003.
- [6] Christiane Fellbaum. 1998. Editor. WORDNET: an electronic lexical database. MIT Press.
- Piek Vossen (ed). EuroWordNet General Document. Version 3, Final, July 19, 1999, <http://www.hum.uva.nl/~ewn>
- [7] Alessandro Lenci, Federica Busa, Nilda Ruimy, Elisabetta Gola, Monica Monachini, Nicoletta Calzolari, Antonio Zampolli, Emilie Guimier, Gaëlle Recourcé, Lee Humphreys, Ursula Von Rekovsky, Antoine Ogonowski, Clare McCauley, Wim Peters, Ivonne Peters, Robert Gaizauskas, Marta Villegas. 2000. SIMPLE Work Package 2 - Linguistic Specifications, Deliverable D2.1. ILC-CNR, Pisa, Italy.
- [8] Piek Vossen (ed). EuroWordNet General Document. Version 3, Final, July 19, 1999, <http://www.hum.uva.nl/~ewn>
- [9] Paul Buitelaar, Thierry Declerck, Anette Frank, Stefania Racioppa, Malte Kiesel, Michael Sintek, Ralf Engel, Massimo Romanelli, Daniel Sonntag, Berenike Loos, Vanessa Micelli, Robert Porzel, Philipp Cimiano LingInfo: Design and Applications of a Model for the Integration of Linguistic Information in Ontologies. In: Proc. of OntoLex06, a Workshop at LREC, Genoa, Italy, May 2006.