Using Information Extraction Rules for Extending Domain Ontologies
- Position Statement for IJCAI-2001 Workshop on Ontology Learning -

Michael Sintek, Markus Junker, Ludger van Elst, and Andreas Abecker
German Research Center for Artificial Intelligence (DFKI)
– Knowledge Management Group –
P.O. Box 2080, D-67608 Kaiserslautern, Germany
Phone: +49 631 205 3470, Fax: +49 631 205 3210
e-Mail (sintek,junker,elst,aabecker)@dfki.de

1 Ontologies in the FRODO Project

In the FRODO project [1] we aim at the development of a "Framework for Distributed Organizational Memories" (OMs). We start with the observation that knowledge and expertise is always heavily distributed in an organization. We accept the fact that this is not an intermediary, imperfect state which should be overcome by a central, ontologically structured information system, but rather a natural and meaningful situation (because during the introduction of OM systems it is normal to start with small, focussed systems which should interoperate later; because much expertise is better to be created, hold, and maintained locally; or because in the case of interorganizational collaborations or virtual teams a deeper integration of information systems cannot be achieved).

Hence, a main goal of the FRODO project is to develop a scalable, extensible OM middleware built for easy integration of new components and linking of collaborating components [2]. FRODO builds upon the KnowMore framework for contextually-aware, ontology-based OMs [3,4], but relaxes some constraints of the original model, especially the idea of a centralized OM using one overall set of organizational ontologies.

Besides the technical provisions for such a distributed, highly dynamic environment, we lay special emphasis on considerations and methods which are necessary to realize such a scenario in industrial practice. In each industrial environment, besides the questions of smooth introduction of new technology regarding human factors and organizational processes, and besides the question of modeling tools and method support for knowledge (in particular ontologies for structuring OMs or parts of OMs) acquisition, at least two other factors are of utmost importance:

One is the predominance of informal, i.e. essentially text-based, representations of knowledge. This is not only just a matter of fact, but really useful, because the cost of formalization is often not in the right relation to the potential benefits such that many informal parts of the scenario are economically reasonable [5]. One implication is that also methods for building formal models must be affordable.

The other is the fact that ontologies are not a stand-alone component built once and then remaining untouched, but a living element in the overall scenario, used for different purposes, communicating with other system parts, and representing knowledge about a continuously changing world [10].

These two assumptions lead to two characteristics of our approach:

- Learning ontological information from text documents should be a main component of the overall scenario. We set the goal already in [3]. In the meanwhile we sketched a method for business-process oriented knowledge modeling in the company, realized as an amalgamation of the CommonKADS [6] and the IDEF5 [7] suites of methods [2]. We build upon the Protégé-2000 knowledge acquisition and modeling tool [8] which we extended already by some modules for modeling, reasoning, and visualization (see [1]). We are currently working on an integration of the MindAccess(r) commercial [9] text analysis workbench which employs a numberof statistical document feature extraction and document analysis functionalities.
- In order to cope with the complexity and dynamics of real-world usage scenarios for ontologies in a distributed OM, we develop a methodological framework for understanding and organizing the roles, responsibilities, rights, and obligations of actors constituting an ontology society in a complex, agent-based OM system [10].

In the IJCAI-01 "Ontology Learning" workshop we would like to discuss primarily an approach for extending the above statistically-oriented learning techniques towards a more knowledge-based one using an ILP (Inductive Logic Programming [11]) algorithm which can use more elaborated document models and can cope with different sources of sophisticated background knowledge.

2 Ontology Learning with Information Extraction Rules

Figure 1 illustrates the overall idea of building ontologies with learned information extraction rules. We start with:

1. An initial, hand-crafted seed ontology of reasonable quality which contains already the relevant types of relationships between ontology concepts in the given domain.
2. An initial set of documents which exemplarily represent (informally) substantial parts of the knowledge represented formally in the seed ontology.

Figure 1: Overall approach for ontology learning with information extraction rules

Now we assume that similar ontological phenomena—e.g. the fact that relationship R holds between concept A and concept B—are expressed in the text in similar ways. Consider, e.g., a medical domain where the fact that Disease A can be treated (this is the Relationship R) with Cure B. Such A-R-B instances of relationship R could, for instance, look like:

- My headache was cured by medication with Aspirin.
- Sue's headache was addressed with acupuncture.
- Cancer can be treated with chemotherapy.
- Cancer is often treated with surgery.

Our main idea is that, (i) given such texts are available which explain the ontological knowledge, and (ii) given these texts are sufficiently similar with respect to the question how similar factual statements are textually represented, it should be possible:

1. To take the pairs of (ontological statement, one or more textual representations) as positive examples for the way how specific ontological statements can be reflected in texts. There are two possibilities to extract such examples:
   - Based on the seed ontology, the system looks up the signature of a certain relation (e.g., R links a Disease with a Cure), searches all occurrences of instances of the concept classes Disease and Cure, respectively, within a certain maximum distance, and regards these co-occurrences as positive examples for relationship R. This approach presupposes that the seed documents have some "definitional" character, like domain specific lexica or textbooks.
   - The user goes through the seed documents with a marker and manually highlights all interesting passages as instances of some relationship. This approach is more work-intensive, but promises faster learning and more precise results. We employed this approach already successfully in an industrial information extraction project [12].
2. Employ a pattern learning algorithm to automatically construct information extraction rules which abstract from the specific examples, thus creating general statements which text patterns are an evidence for a certain ontological relationship. In the example above, such an information extraction rule could have the form:

   > In order to detect an instance of the "Method B is a possible Cure for Disease A" relationship, search for an instance of the concept Disease, look whether there is a synonym of the word (stem) "treat" in a distance of at most two words, search for the word "with" in a distance of at most two words, directly followed by an instance of the concept Cure.

   In order to learn such information extraction rules, we need some prerequisites:

   (a) A sufficiently detailed representation of documents (in particular, including word positions, which is not usual in conventional, vector-based learning algorithms, WordNet-synsets, and part-of-speech tagging).
   (b) A sufficiently powerful representation formalism for extraction patterns.
   (c) A learning algorithm which has direct access to background knowledge sources, like the already available seed ontology containing statements about known concept instances, or like the WordNet database of lexical knowledge linking words to their synonyms sets, giving access to sub- and superclasses of synonym sets, etc.

   In [13,14] we present an ILP-like rule learner specifically adapted to the task of pattern-based text classification (which can be solved with the same methods as the information extraction task used in the ontology learning application) which fulfills these requirements. In particualar, this rule learner relies on a document representation in which the order of words is preserved. Thus, learned text patterns can test on the order and distance of specific words. In [16] it is shown how its implementation concepts can be mapped to standard ILP approaches, which shows how its expressive power with respect to pattern representation can even be extended towards full LP formalisms including recursive rules. In [15] we elaborate a bit on the integration of background knowledge sources, especially WordNet.
3. Apply these learned information extraction rules to other, new text documents to discover new or not yet formalized instances of relationship R in the given application domain.

3 Status

The algorithm described has not yet been implemented and tested. However, all required prerequisites are available as described above and in [13,14,15,16]. Further, we are in contact with several application projects (in the nuclear and the chemical industry) in order to get significant test data. A critical factor for the success of the approach will be the question of how typical the textual representations of specific (kinds of) statements will be in the seed documents.

Compared to other ontology learning approaches it should be noted that our technique is not restricted to learning taxonomic relationships, but arbitrary relationships in an application domain. We expect that, in contrast to more statistically oriented approaches, which tend to result in too many candidate results (because of many possibly relevant word co-occurences), our approach needs more input and assumes

more prerequisites, but found relationship candidates will be correct with a higher probability.

References

1. FRODO project homepage: http://www.dfki.uni-kl.de/frodo/
2. Abecker, A., Bernardi, A., van Elst, L., Lauer, A., Maus, H., Schwarz, S., and Sintek, M. (2001). FRODO: A Framework for Distributed Organizations - Milestone M1: Requirements Analysis and System Architecture. DFKI Document D-01-01. In preparation. Partially in German.
3. Abecker, A., Bernardi, A., Hinkelmann, K., Kühn, O., and Sintek, M. (1998). Towards a Technology for Organizational Memories. IEEE Intelligent Systems, 13(3), May/June.
4. Abecker, A., Bernardi, A., Hinkelmann, K., Kühn, O., and Sintek, M. (2000). Context-Aware, Proactive Delivery of Task-Specific Knowledge: The KnowMore Project. International Journal on Information System Frontiers, Kluwer, 2(3/4).
5. Buckingham Shum, S. (1997). Balancing Formality with Informality: User-Centred Requirements for Knowledge Management Technologies. AIKM'97: AAAI Spring Symposium on Artificial Intelligence in Knowledge Management, Stanford University, Palo Alto, CA. AAAI Press.
6. Schreiber, G., Akkermans, H., Anjeiwerden, A., de Hoog, R., Shadbolt, N., van de Velde, W., and Wielinga, B. (1999). Knowledge Engineering and Management: The CommonKADS Methodology. MIT Press.
7. Information Integration for Concurrent Engineering (1994). IDEF5 Method Report. URL: http://www.idef.com/ .
8. Grosso, W.E., Eriksson, H., Fergerson, R.W., Gennari, J.H., Tu, S.W., and Musen, M.A. (1999). Knowledge Modeling at the Millennium (The Design and Evolution of Protege-2000). SMI-1999-0801. Stanford Medical Lab. URL: protege.stanford.edu
9. MindAccess product description (2000). Insiders information management GmbH, Kaiserslautern. URL: http://www.im-insiders.de/html/infomaterial.html. In German.
10. van Elst, L. and Abecker, A. (2001). Ontology-Related Services in Agent-Based Distributed Information Infrastructures. Submitted to: SEKE'01, The Thirteenth International Conference on Software Engineering & Knowledge Engineering, June 13-15, 2001, Buenos Aires - Argentina
11. Lavrac, N. and Dzeroski, S. (1994). Inductive Logic Programming: Techniques and Applications. Chichester, UK: Ellis Horwood.
12. ANNOCLASS project description. URL: http://www.dfki.de/pas/f2w.cgi?daimc/annoclass-e
13. Junker, M. (2000). Heuristisches Lernen von Regeln für die Textkategorisierung. Dissertation. Fachbereich Informatik. Universität Kaiserslautern. In German.
14. Junker, M. and Abecker, A. (1998). Learning Complex Pattern for Document Categorization. In: AAAI-98/ICML Workshop on Learning for Text Categorization. Madison, Wisconsin, USA.
15. Junker, M. and Abecker, A. (1997). Exploiting Thesaurus Knowledge in Rule Induction for Text Classification. In: RANLP'97 - Recent Advances in NLP, pp. 202-207, Tzigov Chark, Bulgaria.
16. Junker, M., Sintek, M., and Rinck, M. (2000). Learning for Text Categorization and Information Extraction with ILP. In Learning Language in Logic, Springer, LNCS 1925.