User-Centred Ontology Learning for Knowledge Management

Christopher Brewster, Fabio Ciravegna and Yorick Wilks

Department of Computer Science, University of Sheffield
Regent Court, 211 Portobello Street, Sheffield, UK
{C.Brewster|F.Ciravegna|Y.Wilks}@dcs.shef.ac.uk

Abstract. Automatic ontology building is a vital issue in many fields where they are currently built manually. This paper presents a user-centred methodology for ontology construction based on the use of Machine Learning and Natural Language Processing. In our approach, the user selects a corpus of texts and sketches a preliminary ontology (or selects an existing one) for a domain with a preliminary vocabulary associated to the elements in the ontology (lexicalisations). Examples of sentences involving such lexicalisation (e.g. ISA relation) in the corpus are automatically retrieved by the system. Retrieved examples are validated by the user and used by an adaptive Information Extraction system to generate patterns that discover other lexicalisations of the same objects in the ontology, possibly identifying new concepts or relations. New instances are added to the existing ontology or used to tune it. This process is repeated until a satisfactory ontology is obtained. The methodology largely automates the ontology construction process and the output is an ontology with an associated trained leaner to be used for further ontology modifications.

1. Introduction

The importance of ontologies is widely accepted in a number of domains including the Semantic Web, Knowledge Management and electronic commerce [1][2]. They provide a means to structure and model the concepts shared by a group of people concerning a specific domain. While a great deal of effort is going into planning the use of ontologies, much less has been achieved in automating their construction: in making feasible a computational process of knowledge capture.

Ontologies traditionally are built entirely by hand and the source of information for these knowledge structures is usually introspection or protocol analysis [3]. In this context, the automation of the process of knowledge capture is still in its infancy. The process of knowledge capture or ontology construction can be analyzed as involving three major steps: first, the construction of a concept hierarchy; secondly, the labelling of relations between concepts, and thirdly, the association of content with each node in the ontology [4]. The dynamic nature of human knowledge makes an automatic system that can be trained on real data (i.e. texts) an imperative.

In the past, a number of researchers have proposed methods for creating conceptual hierarchies or taxonomies of terms from processing texts by applying methods from Information Retrieval (term distributions in documents) and Information Theory (mu-

tual information) [2]. It is relatively easy to show that two terms are associated in some manner or to some degree of strength [5][6]). It is possible also to group terms into hierarchical structures of varying degree of coherence [7][8]. However, the most significant challenge, which has not been resolved, is to be able to label the nature of the relationship between the terms [9]. Only if relations are explicit can an ontology be used with problem solving methods (PSMs) [10] i.e. for some form of logical inference.

A set of "lexico-syntactic patterns" which would identify specific ontological relations was proposed by Hearst [11] and later implemented by Morin [12] but with repeated user intervention. Building on this work, and in order to build ontologies for real world applications in Knowledge Management (KM), we propose a methodology based on a co-operative model of user and system interaction. The model is based on the integration of Natural Language Processing (NLP) techniques (especially Information Extraction from text - IE) with user input, so as to limit the user's effort and yet obtain the most accurate possible ontology. Our objective is to make as effective as possible the user's input to the system without expecting any understanding of the nature of 'lexico-syntactic patterns'. The rest of this paper is organized as follows: Section 2 describes the characteristics of the user and the system. We present the major steps in the learning process in Section 3 and an overview of the system interface and learning engine in Section 4. The paper finishes with a description of future work and a conclusion.

2. Building Ontologies for Knowledge Management

We need to have a greater understanding of the qualities and characteristics of the user, who wishes to build an ontology, and the system and its potential capabilities.

User Characteristics. The system we are proposing is developed for the specific context and needs of KM and this implies users with specific characteristics. They are assumed not to have any specialised knowledge but we do assume that they are able to a) draft an ontology, or select or reuse an existing one, and provide this as input to the system; b) validate sentences which are exemplars of a particular relation between two terms; c) name/label a relation exemplified in a particular sentence, and to recognise when they encounter further instances of such a relation. Such characteristics are not specific of KM only, but of a set of users in different fields: for example Semantic Web users tend to have the same profile.

The Characteristics of the System. These are to some extent the characteristics of computer systems in general, but here we focus on those of a combined NLP/IE system. In general they are able to a) analyse large quantities of texts at speeds which often approximate real time; b) find regularities and identify all occurrences of a given regularity; c) cluster words and other patterns into groups; d) establish that a relationship exists between any given term *x* and another term *y*.

These characteristics have already revolutionised lexicography and should have a similar effect on ontology construction and knowledge capture. The ability to find regularities is particularly significant in view of the large quantities of data involved.

3. User - Centred Pattern Learning

The learning process is divided in two stages: the system first attempts to learn about the ISA/hyponymy relations between concepts, and once these have been established (via the steps below) the skeletal ontology is presented to the user who may select further relations to learn. Each of the two stages consists of three steps: bootstrapping, pattern learning and user validation, and cleanup.

Bootstrapping. The bootstrapping process involves the user specifying a corpus of texts, and a seed ontology. The draft ontology must be associated with a small thesaurus of words, i.e. the user must indicate at least one term that lexicalises each concept in the hierarchy.

Pattern Learning & User Validation. Words in the thesaurus are used by the system to retrieve a first set of examples of the lexicalisation of the relations among concepts in the corpus. These are then presented to the user for validation. The learner then uses the positive examples to induce generic patterns able to discriminate between them and the negative ones. Pattern are generalised in order to find new (positive) examples of the same relation in the corpus. These are presented to the user for validation, and user feedback is used to refine the patterns or to derive additional ones. The process terminates when the user feels that the system has learned to spot the target relations correctly. The final patterns are then applied on the whole corpus and the ontology is presented to the user for cleanup.

Cleanup. This step helps the user make the ontology developed by the system coherent. First, users can visualise the results and edit the ontologies directly. They may want to collapse nodes, establish that two nodes are not separate concepts but synonyms, split nodes or move the hierarchical positioning of nodes with respect to each other. Also, the user may wish to 1) add further relations to a specific node; 2) ask the learner to find all relations between two given nodes; 3) refine/label relations discovered in the between given nodes. Corrections are returned back to the IE system for retraining.

This methodology focuses the expensive user activity on sketching the initial ontology, validating textual examples and the final ontology, while the system performs the tedious activity of searching a large corpus for knowledge discovery. Moreover, the output of the process is not only an ontology, but also a system trained to rebuild and eventually retune the ontology, as the learner adapts by means of the user feedback. This simplifies ontology maintenance, a major problem in ontology-based methodologies.

4. Adaptiva

Adaptiva is a system implementing the methodology above that has been developed as part of the AKT project [15]. The ontology learning process starts with the definition of the draft ontology, which is imported into the system's internal format by using a converter. **Adaptiva** is based on GATE [14] which provides facilities for corpus management. Lexicalisation of concepts and relations in the ontology are used to retrieve the first set of examples in the corpus. Such examples are presented to the user

for validation by using a simple interface shown in Figure 1, thus specifying whether the sentence presented is a positive, a negative or an irrelevant example.

The actual complete interface consists of three panes which present i) the examples still to be classified, ii) the examples classified as positive, and iii) those classified as negative. As each example is validated, the user checks one of the two check boxes or leaves the example alone (e.g. because it is too difficult or thought to be irrelevant). According to which box is checked the example moves to the positive or negative pane, thereby allowing the user to revise their decision.

Name of Relation	*Exemplar sentence*	*Positive Example*	*Negative example*
ISA	…countries such as England, France and Italy….	☑	☐

Fig. 1. The interface for user validation

The outcome of the validation process is used by a pattern learner, which in our case is Amilcare (cf. below). Once the learning process is completed, the induced patterns are applied to unseen corpus and new examples are returned for further validation by the user. This iterative process may continue until the user is satisfied that a high proportion of exemplars is correctly classified automatically by the system.

Using a learning algorithm. The methodology described above is generic in that it is not tied to one specific Machine Learning algorithm or approach. The precise methodology is irrelevant from the user's perspective. In **Adaptiva**, we have integrated Amilcare [13], a tool for adaptive Information Extraction from text (IE) designed for supporting active annotation of documents for the Semantic Web.

Using Amilcare, positive and negative examples are transformed into a training corpus where XML annotations are used to identify the occurrence of relations in positive examples. The learner is then launched and patterns are induced and generalised. After testing, the best, most generic, patterns are retained and are then applied to the unseen corpus to retrieve other examples. From Amilcare's point of view the task of ontology learning is transformed into a task of text annotation: the examples are transformed into annotations and annotations are used to learn how to reproduce such annotations.

5. Conclusions and Future Work

We have presented a novel method of user-system interaction for the purposes of ontology building, specifically in the context of knowledge management. This work implements to a larger degree the ideas first proposed by Hearst and built on by Morin. The advantage of our methodology with respect to the previous works is that it does not require any ability to define lexico-semantic patterns. The only knowledge needed is the ability to sketch an ontology and to validate examples, characteristics that are common to users in many application domains. We believe that this is a new direction for user-centred ontology building that could have considerably impact on the way in ontologies are built for real world applications. Future work will concern the evaluation of qualitative and ergonomic aspects so as to establish what the benefits are, and to what degree and how the system can be further improved for the user. It is difficult

to benchmark complex systems such the one presented above, but we are developing criteria to help determine how the system can be improved [2].

Acknowledgements
This work has been carried on in the framework of the Advanced Knowledge Technologies (AKT) project [19], an Interdisciplinary Research Collaboration (IRC) sponsored by the UK Engineering and Physical Sciences Research Council (grant GR/N15764/01). AKT involves the Universities of Aberdeen, Edinburgh, Sheffield, Southampton and the Open University. Its objectives are to develop technologies to cope with the six main challenges of knowledge management: acquisition, modelling, retrieval/extraction, reuse, publication and maintenance. Amilcare is based on GATE [18]. Thanks to the GATE group at the University of Sheffield for providing the Annie system and for help in integrating Amilcare and Annie.

References

1. Fensel, D., F. van Harmelen, I. Horrocks, D.L. McGuiness, P.F. Patel-Schneider, (2001) OIL: An Ontology Infrastructure for the Semantic Web, *IEEE Intelligent Systems* (16)
2. Brewster, C. (2002) Techniques for Automated Taxonomy Building: Towards Ontologies for Knowledge Management, in *Proceeding of the 5th Annual CLUK Research Colloquium*, Leeds
3. Ericsson, K. A. & H. A. Simon, (1984) *Protocol Analysis: verbal reports as data.* Cambridge, Mass.: MIT Press
4. Brewster, C., F. Ciravegna, and Y. Wilks, (2001) Knowledge Acquisition for Knowledge Management: Position Paper, in *Proceeding of the IJCAI-2001 Workshop on Ontology Learning* held in conjuction with the 17th International Conference on Artificial Intelligence (IJCAI-01), Seattle, August, 2001
5. Grefenstette, Gregory, (1994), *Explorations in Automatic Thesaurus Discovery*, Amsterdam: Kluwer.
6. Scott, Michael, (1998) Focusing on the Text and Its Key Words, *TALC 98 Proceedings*, Oxford, Humanities Computing Unit, Oxford University.
7. Brown, Peter F.,Vincent J. Della Pietra, Petere V. DeSouza, Jenefer C. Lai, Robert L. Mercer, (1992) Class-based n-gram models of natural language, *Comp. Linguistics*, (18)
8. Sanderson, Mark, and Bruce Croft, (1999) Deriving concept hierarchies from text, in *Proceedings of the 22nd ACM SIGIR Conference*
9. Wilks, Y. (2000) Artificial Intelligence and Information Retrieval, in *Proceedings of the IEE Symposium on Information Retrieval and Artificial Intelligence*, Glasgow**.**
10. Gomez-Perez, A., (1999) Evaluation of Taxonomic Knowledge in Ontologies and Knowledge Bases, in *Proceedings of the 12th Banff Knowledge Acquisition for Knowledge-Based Systems Workshop*, Banff, Alberta, Canada
11. Hearst, M.A., (1992) Automatic Acquisition of Hyponyms from Large Text Corpora, in *Proceedings of COLING 92*, Nantes
12. Morin, Emmanuel, (1999b), Using Lexico-Syntactic patterns to Extract Semantic Relations between Terms from Technical Corpus, in *Proceedings of TKE 99*, Innsbruck, Austria,
13. Ciravegna, Fabio, Alexiei Dingli, Daniela Petrelli, (2002) Document Annotation via Adaptive Information Extraction, *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* August 11-15, 2002, Tampere, Finland.
14. GATE, http://www.gate.ac.uk/
15. Advanced Knowledge Technologies, http://www.aktors.org