Knowledge Acquisition for Knowledge Management: Position Paper

Christopher Brewster, Fabio Ciravegna and Yorick Wilks

Department of Computer Science, The University of Sheffield,
Regent Court, 211 Portobello Street, S1 4DP, Sheffield, UK
{C.Brewster|F.Ciravegna|Y.Wilks}@dcs.shef.ac.uk

1. Introduction

Knowledge is considered to be "the information needed to make business decisions"[1], and so knowledge management is the "essential ingredient of success" for 95 per cent of CEOs [Manchester 1999]. A company's value depends increasingly on "intangible assets"[2] which exist in the minds of employees, in databases, in files and in a myriad documents. Knowledge management technologies capture this intangible element in an organisation; and make it universally available. The most widely used method of mapping the knowledge of a domain is to use an ontology describing such a domain. Ontologies can act as an index to the memory of an organisation and facilitate semantic searches and the retrieval of knowledge from the corporate memory as it is embodied in documents and other archives. There are many real-world examples where the utility of ontologies as maps or models of specific domains has been repeatedly proven.

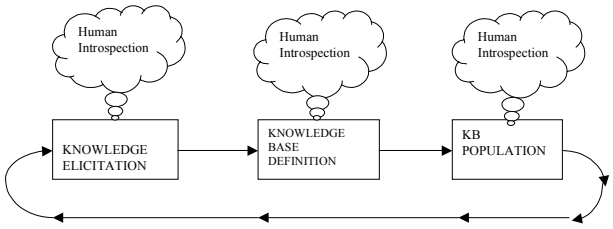

Figure 1 A basic knowledge base development cycle (current reality)

For knowledge to be managed it must first of all be captured or acquired in some useful form, e.g. stored in an ontology. Knowledge acquisition (KA) is a complex process which traditionally is extremely expensive. Yahoo currently employs over 100 people to keep its category hierarchy up to date [Dom 1999]. An appropriate model of knowledge acquisition under the current paradigm would note that it is an entirely manual process. The typical process of KA is outlined in Fig. 1: initially knowledge is elicited either from a set of documents or from an expert of the domain. Then an ontology is defined, finally a set of instances are associated to concepts in the knowledge base. It is widely agreed that the key factors which impede the wider use of ontologies both in research and commercial applications are time, cost and subjectivity. Time and money committed to ontology development is substantial as the years of development at Cyc have shown and the manpower requirements of Yahoo prove. Subjectivity is inevitable, given any method which depends on individual introspection or elicitation for data collection.

Although efforts exist to automate KA most of these are isolated to specific aspects of the construction cycle, little effort has been spent on creating a comprehensive set of tools for KA. In this paper, we propose a set of techniques to largely automate the process of KA, by using technologies based on Information Extraction (IE), Information Retrieval and Natural Language Processing. We aim to reduce all the impeding factors mention above and thereby contribute to the wider utility of the knowledge management tools. In particular we intend to reduce the introspection of knowledge engineers or the extended elicitations of knowledge from experts by extensive textual analysis using a variety of methods and tools, as texts are largely available and in them – we believe – lies most of an organization's memory.

2. Ontology Learning

Ontology construction aims to capture knowledge in a usable format. The nature and granularity of the ontology depends on its eventual use; typically IS-A hierarchies are central but other emphases are often encountered.

The process of ontology construction may be divided into three stages the first two of which contribute to the learning of the ontology structure and the third is used to populate

[1] Philip Crawford, European Vice-President of Oracle Corp. cited by [Philip 1999]
[2] A term coined by the industry consultant Karl-Erik Sveiby

the knowledge base with instances. These stages are illustrated in the rest of this section.

2.1 Taxonomy construction

We propose to introduce automation in the stage of taxonomy construction mainly in order to eliminate or reduce the need for extensive elicitation of data. In the literature approaches to construction of taxonomies of concepts have been proposed [Brown *et al.* 1992, McMahon and Smith 1996, Sanderson and Croft 1999]. Such approaches either use a large collection of documents as their sole data source, or they can attempt to use existing concepts to extend the taxonomy [Agirre *et al.*2000, Scott 1998]. We intend to develop a semi-automatic method that, starting from a seed ontology sketched by the user, produces the final ontology via a cycle of refinements by eliciting knowledge from a collection of texts. In this approach the role of the user should only be that of proposing an initial ontology and validate/change the different versions proposed by the system. We believe an ontology construction method should a) permit multiple placement of terms in the structure, b) allow rapid recalculation of the structure, c) provide monothetic labels for nodes, d) allow the input of seed ontologies for further expansion.
We intend to integrate a methodology for automatic hierarchy definition (such [Sanderson and Croft 1999]) with a method for the identification of terms related to a concept in a hierarchy (such as [Scott 1998]).

The advantage of this integration is that as knowledge is continually changing, we can reconstruct an appropriate domain specific ontology very rapidly. This does not preclude incorporating an existing ontology and using the tools to extend and update it on the basis of appropriate texts. Finally an ontology defined in this way has the particular advantage that it overcomes the well-known 'Tennis problem' associated with many predefined ontologies such as WordNet, i.e where terms closely related in a given domain are structurally very distant such as *ball* and *court*.

In addition we intend to employ classic Information Extraction techniques such as Sheffield's named entity recognition system [Humphreys 1998] in order to preprocess the text, as the identification of complex terms such as proper names, dates, numbers, etc, allows to reduce data sparseness in learning [Ciravegna 2000].

We plan to introduce many cycles of ontology learning and validation. At each stage the defined ontology can be: i) validated/corrected by a user/expert; ii) used to retrieve a larger set of appropriate documents to be used for further refinement [Järvelin and Kekäläinen 2000]; iii) passed on to the next development stage below.

2.2 Learning Other Relations

This stage proceeds to build on the skeletal ontology in order to specify, as much as possible without human intervention, relations among concepts in the ontology, other than ISAs. In order to flesh out the concept relations, we need to identify relations such as synonymy, meronymy, antonymy and other relations. We plan to integrate a variety of methods from the literature, e.g. by using recurrences in verb subcategorisation as a symptom of general relations [Basili *et al.* 1998], by using Morin's user-guided approach to identify the correct lexico/syntactic environment [Morin 1999], and by using methods such as [Hays 1997] to locate specific cases of synonymy.

3. Populating the Knowledge Base

Once the ontology has been learnt, there is the problem of retrieving instances in order to populate the resulting knowledge base. This is a key issue, in order to use the ontology as index to the organisation's memory, for example by allowing semantic searches and the retrieval of knowledge from the corporate memory. In many environments KB population is performed manually by a user via instance identification in a text corpus. We plan to automate this process as much as possible by using a combination of text classification (TC) (e.g. [Ciravegna et al 1999]) and Adaptive Information Extraction ([Ciravegna 2001]). Text classification is useful in order to identify the scenario to apply to a specific set of texts, while IE will identify (i.e. index) the instances in the texts. Both TC and IE should be adaptive, as it is generally not possible to ask a user to develop rules himself for each scenario/application. Again, automating the process will both reduce the cost of instance identification and the subjectivity involved in the human identification.

4. Conclusion and future work

Knowledge is only of value when it can be used effectively and efficiently. The management of knowledge is a key element in extracting its value. In this position paper we have outlined how we are addressing the issue of automating the Knowledge Acquisition process in order to reduce both required time and cost of KA, and subjectivity in the resulting ontology. Overall we believe, this will make knowledge management not only more acceptable in a commercial environment but also contribute to the overall productivity of the economy.

The work outlined above is being undertaken by the University of Sheffield in the context of AKT (Advanced Knowledge Technologies, http://www.aktors.org), a multi-million pound, six-year project involving the University of Southampton, the Open University, the University of Edinburgh, the University of Aberdeen, and the University of Sheffield. AKT will extend knowledge management technologies to exploit the potential of the semantic web, covering the use of knowledge over its entire lifecycle, from acquisition to maintenance and deletion. It began in October 2000 and will comprehensively addresses six main challenges, which are fundamental bottlenecks to knowledge management:

- acquisition
- modelling
- retrieval/extraction
- reuse
- publication
- maintenance

The work at Sheffield will on provide a library of Natural Language Processing based tools for different types of knowledge management tasks.

Acknowledgement

This work is supported under the Advanced Knowledge Technologies (AKT) Interdisciplinary Research Collaboration (IRC), which is sponsored by the UK Engineering and Physical Sciences Research Council under grant number GR/N15764/01. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing official policies or endorsements, either express or implied, of the EPSRC or any other member of the AKT IRC.

References

[Agirre *et al.* 2000] Eneko Agirre, Olatz Ansa, E. Hovy, and D. Martínez, Enriching very large ontologies using the WWW, in *Proceedings of the ECAI 2000 workshop "Ontology Learning"* 2000

[Basili *et al.* 1998] Basili, R., R. Catizone, M. Stevenson, P. Velardi, M. Vindigni, Y. Wilks *An Empirical Approach to Lexical Tuning*. In Proc. of the Adapting Lexical and Corpus Resources to Sublanguages and Applications Workshop, held jointly with 1st LREC Granada, Spain, 1998.

[Brown *et al.* 1992] 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, *Computational Linguistics*, 18, 467-479

[Ciravegna *et al.* 1999] Fabio Ciravegna, Alberto Lavelli, Nadia Mana , Luca Gilardoni, Silvia Mazza, Johannes Matiasek, William Black, Fabio Rinaldi, David Mowatt "Classifying Texts Integrating Pattern Matching and Information Extraction" Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence (IJCAI99), Stockholm, August, 1999

[Ciravegna 2001] Fabio Ciravegna "Adaptive Information Extraction from Text by Rule Induction and Generalisation" in Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence (IJCAI 2001), Seattle, August 2001

[Dom 1999] Byron Dom, Automatically finding the best pages on the World Wide Web (CLEVER), In *Search Engines and Beyond: Developing efficient knowledge management systems*, April 19–20 1999, Boston, Mass

[Hays 1997] Paul R. Hays, *Collocational Similarity: Emergent Patterns in Lexical Environments*, Dissertation submitted to the School of English, University of Birmingham

[Hearst 1992] Marti Hearst, Automatic Acquisition of Hyponyms from Large Text Corpora, *COLING 92*, Nantes, 1992

[Humphreys 1999] K. Humphreys, R. Gaizauskas, S. Azzam, C. Huyck, B. Mitchell, H. Cunningham, Y. Wilks. Description of the University of Sheffield LaSIE-II System as used for MUC-7. In *Proceedings of the Seventh Message Understanding Conference (MUC-7)*. Morgan Kaufmann. 1999

[Järvelin and Kekäläinen 2000] Kalervo Järvelin and Jaana Kekäläinen IR evaluation methods for retrieving highly relevant documents. In *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 41-48, Athens, Greece, 24.-28.7.2000,

[Manchester 1999] Philip, Manchester, "Survey – Knowledge Management" *Financial Times*, 28 April, 1999

[McMahon and Smith 1996] McMahon, John G,Francis J. Smith, 1996 Improving Statistical Language Models Performance with Automatically Generated Word Hierarchies, *Computational Linguistics*, 22(2), 217-247, ACL/MIT

[Morin 1999] Emmanuel Morin, Using Lexico-Syntactic patterns to Extract Semantic Relations between Terms from Technical Corpus, *TKE 99*, 268-278, Innsbruck, Austria, 1999.

[Sanderson and Croft 1999] Mark Sanderson and Bruce Croft, Deriving concept hierarchies from text, in *Proceedings of the 22nd ACM SIGIR Conference*, 206-213, 1999

[Scott 1998] Mike Scott, Focusing on the Text and Its Key Words, *TALC 98 Proceedings*, Oxford, Humanities Computing Unit, Oxford University, 1998.