Techniques for Automated Taxonomy Building: Towards Ontologies for Knowledge Management

Christopher BREWSTER
Department of Computer Science,
University of Sheffield,
Sheffield, United Kingdom, S1 4DP
C.Brewster@dcs.shef.ac.uk

Abstract

Ontologies have become widely accepted as the main method for representing knowledge in Knowledge Management (KM) applications. Given the continuous and rapid change and dynamic nature of knowledge in all fields, automated methods for constructing ontologies are of great importance. All ontologies or taxonomies currently in use have been hand built and require considerable manpower to keep up to date. Taxonomies are less logically rigorous than ontologies, and in this paper we consider the requirements for a system which automatically constructed taxonomies. There are a number of potentially useful methods for constructing hierarchically organised concepts from a collection of texts and there are a number of automatic methods which permit one to associate one word with another. The important issue for the successful development of this research area is to identify techniques for labelling the relation between two candidate terms, if one exists. We consider a number of possible approaches and argue that the majority are unsuitable for our requirements.

1. The Need for Ontologies and Taxonomies

Artificial intelligence has for decades confronted the issue of 'knowledge acquisition' and struggled to move from representations of toy worlds to larger more realistic representations of human knowledge. One current form of this struggle has crystallised in the commercial needs of Knowledge Management (KM). In the modern 'knowledge-based' economy, a company's value depends increasingly on "intangible assets" 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. Repeated research has shown their usefulness [include the Mädche refs], especially for specific domains (Järvelin & Kekäläinen 2000). For example, in order to successfully manage a complex knowledge network of experts, the Minneapolis company Teltech has developed an ontology of over 30,000 terms describing its domain of expertise (Davenport 1998).There are many real-world examples where the utility of ontologies as maps or models of specific domains has been repeatedly proven (Fensel *et al.* 2001).

The use of ontologies in KM and in the Semantic Web have had particular momentum, especially the latter as they are central to Tim Berners-Lee's vision of the Semantic Web (Berners-Lee *et al.* 2001) and the ability of agents to perform "sophisticated tasks for users". Whatever the discipline, however, existing work on the construction of ontologies has concentrated on the formal properties and characteristics that an ontology should have in order to be useful (Gomez-Perez 1999, Guarino REFS) rather than the practical aspects of constructing one i.e. to reduce the enormous manual effort involved.

The rest of this paper is organised as follows. In Part 2, we discuss some of the problems associ-

ated with knowledge acquisition as seen from the prism of building ontologies and taxonomies. In Part 3, we will consider a number of methodological criteria which arise from the context of trying to build taxonomies for knowledge management and how they affect our choice of algorithms and methods. In Parts 4 and 5, we briefly discuss some of the major techniques in the literature concerning building hierarchies, and identifying relations between terms in texts, respectively.

2. The Problem with Knowledge Acquisition

Knowledge, is is widely assumed, can be codified in an ontology. An ontology has been define by Gruber (1993) as a "formal explicit specification of a shared conceptualisation" and this has been widely cited with approval (Fensel *et al.* 2001). Berners-Lee says "an ontology is a document or file that formally defines the relations among terms. The most typical kind of ontology for the Web has a taxonomy and a set of inference rules" (Berners-Lee *et al.* 2001). We see ontologies as lying on a continuum reflecting the degree of logical rigour applied in their construction. At the one extreme lie ontologies which purport to be entirely explicit in the sense that logical inferences can, in principle, be easily calculated over these structures. At the other extreme we could place pathfinder networks (Schwaneveldt 1990) or even 'mind-maps' (Buzan 1993), which essentially involve considerable human interpretation to be said to represent 'knowledge' of any form. Somewhere in between lie taxonomies and browsable hierarchies which are clearly less rigorous than a fully specified ontology. Our interest in this paper lies in the construction of taxonomies and browsable hierarchies because we believe that it is more feasible to construct these automatically or semi-automatically than fully-fledged ontologies. Gomez-Perez (1999), for example, presents very strict criteria for ontology construction concerning *consistency, completeness* and *conciseness* which may be achievable in a specific sub-domain (she discusses the 'Standard Units' ontology) but can only be idealised objectives when dealing with wider knowledge areas. This is entirely parallel with the art of lexicography, which also aspires to exactly the same ideals, but which any experience lexicographers knows are just that: 'ideals'.

One of the major problems in this field is that it is a common conception among authors working with ontologies to assume that ordinary users will be willing to contribute to the building of a formal ontology. Thus for example, Stutt and Motta presents an imaginary scenario where an archaeologist marks up his text with 'various' ontologies and furthermore not finding the Problem Solving Methods (PSMs) associated with the ontologies adequate, adds to the set of existing PSMs (Stutt and Motta 2000:218). This is entirely unrealistic because there is no motivation for archaeologists to burden themselves with this kind of extra task. Similar conclusion have been drawn in industry. It was assumed given the existence of a taxonomy or ontology, authors will be willing to tag their own work in an appropriate manner but the experience of both librarians historically and more recently companies like ICL and Montgomery Watson is that authors tag inadequately or inappropriately their own work (Gilchrist and Kibby 2000).

Currently ontologies and taxonomies are all hand-built. Whether we consider the general browsing hierarchies of Yahoo or Northern Lights at one extreme or the narrow scientific ontology developed by the partners of the Gene Ontology project (http://www.geneontology.org/), these data structures are built by manual labour. Yahoo is reputed to employ over a one hundred people to keep its taxonomy up to date (Dom 1999). Although considerable use is made of taxonomies in industry, it is clear from a number of sources that they are all the result of manual effort both in construction and maintenance. Consider this extract for example from a recent job advertisement on Dice.com:

> Duties: The Ontology Manager will hold a key role in meeting customer demands by maintaining the master ontology that organized content for the eBusiness initiatives. This individual will ensure the data is organized to facilitate rapid product selection ….

A typical example is that of Arthur Andersen who have recently constructed a company wide taxonomy entirely by hand. Their view of the matter is that there is no alternative because the categories used come from the nature of the business rather than the content of the docu-

ments. This is paralleled by the attitude of the British Council's information department who view that the optimum balance between human and computer, in this area, is 85:15 in favour of humans. Not all companies perceive human input as so sacrosanct; Braun GmbH for example would appreciate a tool for taxonomy creation and automatic keyword identification (Gilchrist and Kibby 2000:34). One of the earliest exponents of knowledge management, PricewaterhouseCoopers consider that "the computer can enable activity, but knowledge management is fundamentally to do with people (ibid.:118).

One manner in which certain companies reduce the manual effort involved is by using ready-made taxonomies provided by others. An example of this is Braun GmbH whose internal taxonomy is based on the (hand-built) resources provided by FIZ Technik (a technical thesaurus) and MESH (the medical subject headings provided by the US Library of Medicine). Nonetheless about 20% of the vocabulary in the taxonomy is generated internally to the company. Another example is the case of GlaxoWellcome (now GSK) who have integrated three legacy taxonomies derived from company mergers using a tool called Thesaurus Manager developed by the company in collaboration with Cycorp, Inc.

There are major problems with the construction and maintenance of ontologies and taxonomies. First, there is the high initial cost in terms of human labour in performing the editorial task of writing the taxonomy and maintaining it. In fact, this consists of two tasks. One is the construction of the actual taxonomy and the other is associating specific content with a particular node in the taxonomy. For example, in Yahoo or the Open Directory (www.dmoz.org), there is the actual hierarchy of categories and then there are specific web sites which are associated with a particular category. Secondly, the knowledge which the taxonomy attempts to capture is in constant flux, it is changing and developing continuously. This means that if the taxonomy is built by hand, like a dictionary, it is out of date on the day of publication. Thirdly, taxonomies need to be very domain specific. Particular subject areas whether in the academic or business world have their own vocabulary and technical terminology, thus making a general ontology/taxonomy inappropriate without considerable pruning and editing. Fourthly, taxonomies reflect a particular perspective on the world, the perspective of the individuals or organisation which builds them. For example, a consulting firm has in its internal taxonomy the category 'business opportunity' but what artefacts fall within this category is a function of both the nature of the business and the insights the consultants have themselves. Fifth, and this is an extension of the previous issue, often the categories in a taxonomy are human constructs, abstractions reflecting a particular understanding. Thus a category like 'business opportunity' or even 'nouns' is an abstraction derived from an analytical framework and not inherent in the data itself. Finally, the fact remains that while an ontology is supposed to be a "shared conceptualisation", it is often very difficult for human beings to agree on a particular manner to categorise the world.

Given these problems there are two possible conclusions. The first three points indicate the need for maximally automated systems which reduce the manual labour involved and make it feasible to keep a taxonomy up to date. The last three points would seem to indicate that the task is not feasible or at best irrelevant. However, we have argued elsewhere for a model of ontology construction involving the judicious integration of automated methods with manual validation (Brewster *et al.* 2000), and this we believe is the direction to take.

2.1 Protocols, Introspection and Textual Data

There are two traditional methods in KM for the acquisition of knowledge whether it is used to construct an ontology or some other from of knowledge base. The one is protocol analysis (Ericsson & Simon 1984) involving the use of structured interviews of experts in a particular domain, asking them to describe their thought process as they work and the knowledge used to make decisions or arrive at conclusions. The other is human introspection which is widely used for example in the construction of a large number of ontologies available at the Stanford Ontology Server. A parallel can be drawn with linguistics and lexicography. Traditionally in linguistics two approaches were used to write a dictionary. One, characteristic of field linguists

and used when the language was obscure or entirely unknown, involved elicitation i.e. interviews with native informants. This is parallel to a protocol analysis approach. The other, characteristic of lexicographers and used for dictionaries of well-known languages, involved using everyone else's previous dictionaries and ones own introspection. These were the methods used for most dictionary production until the late 1980's. However, under the influence of the COBUILD initiative (Sinclair 1987), the field switched massively to the use of corpora i.e. large collections of texts either as supplemental data sources or as primary data sources. Even field linguists now make a much greater effort to collect textual artefacts (stories, songs, narratives, etc.) in their work with unknown languages.

In a parallel manner, large collections of texts must represent the primary data source for the construction of ontologies and taxonomies for KM. With the rise of corporate intranets, the increasing use of emails to conduct a large proportion of business activity, and the continuous growth of textual databanks in all professions, it is clear that methods which use texts as their primary data source are the most likely to go at least some of the way towards constructing taxonomies and 'capturing' the knowledge required. Given the observations made above about the unwillingness of individuals to 'add' to a taxonomy, or 'mark-up' their own texts, and given the continuous change and expansion of information in all domains, using texts as the main source of data appears both efficient and inevitable. It is in this context that the focus of this paper will be on methods which can take as input collections of texts in some form or another.

3. Methodological Criteria

In this section we consider a number of criteria to be used when choosing methods which process texts and produce taxonomies or components of taxonomies as their output. Our purpose here is twofold. First, we wish to create a set of criteria in order to help guide the choice of appropriate tools to use in the automatic construction of taxonomies. While there are a large number of methods which might conceivably produce appropriate output, in fact only a subset will actually fulfil these criteria. Secondly, we hope thereby to contribute to a means by which different approaches to constructing taxonomies can be evaluated as there is complete dearth of evaluative measures in this field. Writers on ontology evaluation concentrate on a limited number of criteria which are only appropriate to hand crafted logical objects (Gomez Perez 1999, Guarino & Welty 2000).

3.1 Coherence

A basic criterion is one of coherence, i.e. that the taxonomy generated appears to the user to be a coherent, common sense organisation of concepts or terms. There are, however, many ways in which terms or concepts are associated with one another. The term 'grandmother' is associated in each person's mind with specific images, ideas, concepts and experiences. But these specific cases are not universal even for a subgroup and thus would not make sense to a third party. Coherence is dependant on the terms associated in an ontology and the nature of their association being part of the 'shared conceptualisation' Gruber described.

Here it is important to distinguish linguistic from encyclopaedic coherence. Thus in a thesaurus such as Roget (1852/1982) under a specific category (e.g. *smoothness*) we encounter a collection of synonyms of varying degree of closeness. Here we encounter linguistic coherence in the sense that the grouping 'makes sense' given linguistic criteria. A good example of this Wordnet, which organises a large vocabulary according to a linguistically principled hierarchy. However, it does not provide a useful organisational principle for information retrieval, reasoning or knowledge management in general. It is a linguistic resource much like a dictionary is. In contrast in a taxonomy or browsable hierarchy, what is required is that the under a given category we find concepts or terms which are associated in a sensible manner, in a useful manner for the purpose at hand. Thus in Yahoo under *Education* → *Higher Education* → *Universities* we find a list of universities not a list of synonyms for the concept university. In addition, it is important to avoid the 'tennis problem' which exists in Wordnet where terms associated with the concept 'tennis' are structurally very far away because of the *a priori* design of Wordnet.

Given these observations, the notion of coherence must be understood as being application specific. For our purposes in constructing taxonomies for Knowledge Management, in general, the notion of encyclopaedic coherence is primary while linguistic coherence can only play a secondary role depending on the needs of an application and on the extent to which (for example) a specific term in referred to by a number of other synonymous ones. The hierarchical structures generated must maximally be sensible, useful and representative of the associations and juxtapositions of knowledge which human users actually need and make.

Having made this seemly uncontroversial proposal, it is in fact very difficult to evaluate a taxonomy or hierarchy from this perspective. Given a method, given a specific input and output, there are no widely established criteria for deciding that a particular taxonomy is correct or incorrect, othat one is better than another. While, in fields like information retrieval, we can speak of precision and recall, there are no equivalent measures for an ontology or taxonomy. This is because knowledge is not really a quantitative entity, it is not something that anyone has come up with easy ways to measure (witness the controversies surrounding exams in education). Coherence as conceived here is a qualitative parameter which as yet merely begs the question for its evaluation.

3.2 Multiplicity

By multiplicity, we mean the placement of a term in multiple positions in the taxonomy. The criterion of multiplicity needs to be distinguished from semantic ambiguity. There are clearly a large number of terms which are ambiguous in that the have a number of separate definitions. Obvious examples include terms like *class*, which has an entirely different meaning in the domain of railways, sociology and computer programming. These might be distinguished as in dictionaries by means of a subscript: $class_1$, $class_2$, $class_3$, *etc*. On the other hand, there is often a multiplicity of facets for one single term which justify its multiple placement in a taxonomy or ontology depending on the particular focus of that sub-structure. This is a classic problem in librarianship where a book is often concerned with a multiplicity of topics and the necessity in traditional library classification schemes (Dewey, Library of Congress) to place a book under one class mark (where it will be physically placed) caused much controversy and anxiety. Similarly, many concepts can be placed in different positions in a taxonomy depending on the particular facet of the concept one is interested in or emphasising. Thus, to take a simple example, the term *cat* clearly has its place in a taxonomy of *animals* from a purely zoological perspective. It is also obviously a *pet* but the category of *pets* does not in any way fit in with the classification of animals. Many other obvious examples can be given of this.

This common problem has usually been avoided on the basis of Gale, Church and Yarowsky's claim that there exists "one sense per discourse" (Gale *et al.* 1992) but more recent research has found contrary evidence. Krovetz (1998) has shown that a large proportion of senses of ambiguous words may be found in the same document or discourse. An example of this sort is that in a legal context one may use *support* in the context of children or arguments.

As a consequence, the methods of processing texts that has to be used must allow *cat* to occur both in association with the term *animal* or *mammal* and also in association with *pet*. They must take into account that *support* can occur in different senses in the same context. At a minimum, methodologies which force a unique placement for any given term should be avoided. Better still, we need to identify methods which take into account the different senses of a term.

3.3 Ease of Computation

One of the major issues in Knowledge Management is the maintenance of the knowledge bases constructed. As has already been mentioned, an ontology or taxonomy tends to be out of date as soon as it is published or made available to its intended audience. Furthermore, from the developer's and editor's perspective it is important to have feedback on output from the system as quickly as possible in order to evaluate and validate the results. In many contexts there is a continuous stream of data which must be analysed and where each day or week represents an extraordinary large amount of data whose effects upon the overall ontology cannot be determined *a priori*.

Thus it appears to be very important that the methods chosen do not have great complexity and therefore excessive computational cost. This may appear to be an unimportant issue in this time of immense and cheap computational power but when one realises that some algorithms have a complexity or $O(V^5)$ where V is the size of the vocabulary in the text collection, then it can be seen that this is not an insignificant factor in the selection of appropriate methods. The practical significance of this is that in some application contexts computational complexity needs to be seriously considered. There are of course other contexts where it is much less of a concern (where the quantity of data is limited or possible finite).

3.4 Single labels

Another criterion is that all nodes in a taxonomy or hierarchy need to have single labels. Sanderson and Croft (1999) discuss the difference between polythetic clusters (where members of a cluster have some but not necessarily all of a set of features) and monothetic clusters (where all members of the class are guaranteed to have that one feature). The argue that clusters characterised by one feature are much more easily understood by the user. For example, a well-known approach developed at the Xerox Palo Alto Research Centre was called Scatter/Gather (Cutting *et al.* 1992) where documents would be organised into hierarchies and a set of terms would be extracted from the documents to characterise each cluster. A group of documents might be characterised by the set of terms *{battery California technology mile state recharge impact official cost hour governement}* which while comprehensible is not very easy to use and is discouraging for most users (Sanderson and Croft 1999:1). If Yahoo at every level would label a node by a large collection of terms associated with the topic considerable confusion would be caused. Thus in order to be easy to use, nodes in a taxonomy need single labels even if this is a term composed of more than one word. This does not mean that synonyms are not to be included, but this is different from using a set of disparate terms to characterise a subject area. Synonyms can act as alternative labels for a particular node much as synsets in Wordnet have often a number of synonyms.

Methodologies which produce single labels for a node are to be preferred to ones (such as Scatter/Gather) which produce multiple labels for a node.

3.5 Data Source

The data used by a specific method needs to be of two sorts. First, documents must be used as the primary data source for the reasons discussed above (Section 2.1). Secondly, it should permit the inclusion of existing taxonomy (a 'seed') as a data structure to either revise or build upon as required.

Ontologies and taxonomies are often legacy artefacts in an institution in that they may be the result of years of work and people are loath to abandon them. As mentioned above (Section 2.), often companies merge and as a result two different companies taxonomies need to be merged. These existing data structures need to be maintained subsequently. Furthermore, many institutions view their taxonomy as reflecting their own worldview and wish to impose this particular perspective for the 'top-level'.

Given these constraints, methods need to be used which take as input primarily documents, but which also have the possibility of using an existing taxonomy or ontology as part of its input and to use the documents to propose additions or alterations to the existing taxonomy. This is essential, of course, from the perspective of maintaining a taxonomy. From a practical perspective, given the existence of many taxonomies for one purpose or another, the use of 'seed' taxonomies will be predominant.

4. Constructing Hierarchies

Ontologies and taxonomies are conceived of as hierarchies. More usually they are thought of a trees but in reality they should be viewed as Directed Acyclic Graphs (DAGs) given that a specific term should be able to have more than one parent. The nature of hierarchies is that the more general term is higher (nearer the root) and the more specific is lower down nearer the leaves of the tree. Further more, there are fewer terms higher up, and a greater number of terms lower down. Human beings appear to find such structures particularly easy to understand and well-suited to organising and presenting the structure of knowledge in any domain. Consid-

erable effort has been spent on constructing hierarchies of knowledge probably since Aristotle (4th cent. BC) and certainly since St. Isidore of Seville's *Etymologiarum sive originum libri* (6th cent. AD).

There are a number of methods which organise the vocabulary of a corpus of texts into tree-like structures. One of the best known is to be found in the work of Brown *et al.* (1992) who were attempting to improve language models for speech recognition. This method is based on assigning each vocabulary item to its own class and merging classes where there is minimal loss of mutual information (Church and Hanks 1990, Cover and Thomas 1991). The *order* in which clusters are merged provides a binary tree with intermediate nodes corresponding to groupings of words. Some results are striking: *{mother wife father son husband brother daughter sister boss uncle}* and some not so effective *{ rise focus depend rely concentrate dwell capitalize embark intrude typewriting}* . There are problems with this approach in three areas: ease of construction, labelling, and data source. Computational cost is a problem because their algorithm has a complexity of $O(V^5)$ where V is the size of the vocabulary. This means that in practice the algorithm was applied to the most frequent 5000 or so lexical items. Another more serious issue is that the approach does not provide any means to label the intermediate nodes in the hierarchy generated. Given a node with a class of terms below it, there is no principled way to choose one item to label that class. Finally, this approach does not allow the use of a seed taxonomy on which to build further.

McMahon and Smith (1996) take a closely related approach where the whole vocabulary is assigned to one class and this is divided into two subclasses which maximise mutual information. In order to simplify the process and reduce the computational cost they process only the 500 odd most frequent words, and then deal with the rest of the vocabulary assuming this top level is immutable. The results are impressive with many of the classes appearing to be very coherent. This approach lessens the computational problem but it still cannot provide labels for the nodes generated in the hierarchy.

The Scatter/Gather methodology (Cutting *et al.* 1992), mentioned above, could also be seen as a methodology for constructing hierarchies or concepts. Scatter/Gather has two components: one to cluster documents, and the other to generate a 'cluster digest' i.e. a set of words characterising the cluster of documents. Considerable effort was given by Cutting *et al.* to make the approach efficient and they proposed two different algorithms for clustering documents. While the purpose of this approach is to organise documents, the hierarchical structures generated together with the 'cluster digests' for the documents below each node, make this approach attractive for the generation of taxonomies or ontologies. The main *a priori* problem with this approach, as mentioned above, lies in the fact that each cluster is labelled with a complex set of terms which often are difficult to understand.

One method which avoids the problem of labelling (either too many terms, or none at all) is that presented by Sanderson and Croft (1999). They use a document-based notion of subsumption, where "*x* subsumes *y* if the documents which *y* occurs in are a subset of the documents which *x* occurs in" (This is not to be confused with the traditional notion of subsumption which refers to the IS-A/hyponym relation). More specifically, their approach uses a query term (or terms) to select a set of documents in a corpus, the terms in those documents are identified and then the subsumption relation is calculated between each. The pairs of subsumed terms are then organised into a hierarchy.

This approach fulfils all the requirements described above except for one which is that of coherence. It allows terms to be subsumed by more than one term, it is relatively easy to compute, it provides single terms as labels for nodes, and it is not difficult to imagine how to use the nodes in an existing hierarchy as input for the creation of further sub-hierarchies. The problem of coherence exists because the Sanderson and Croft approach assumes that if X subsumes Y, then X will always be found in a superset of the files Y is found in. This makes sense when one is dealing with the middle level of a taxonomy i.e. from basic terms down to specialised terms. However, often more general terms are used less frequently than the basic level term. The expression 'basic level' is used very loosely here to describe the everyday genus term level like *dog, tree, flower*. Terms like *mammal* are less fre-

quent in most texts than *dog*. For example, in the Nature Corpus[1], the term *basalt* occurs in 159 files, term *rock* in over 800 files, but *igneous rock* in only 29 files. The common sense hierarchical structure would be as in Fig. 1, but

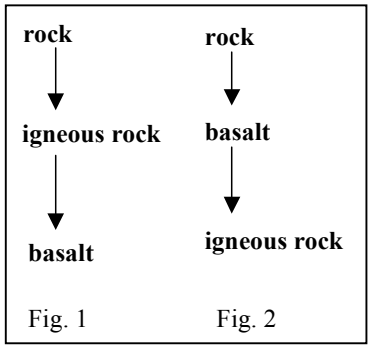

Fig. 1 Fig. 2

Sanderson and Croft's approach would predict a structure as in Fig. 2.

Thus it would appear that the Sanderson and Croft approach may be useful under certain circumstances but it does not provide clearly coherent output.

5. Labelling Relations

The key problem in constructing taxonomies or ontologies lies not in constructing the hierarchy but, assuming that two terms exist, determining what the nature of the relation is between them. There are a large number of methods for identifying the fact that term X and term Y are associated together (Grefenstette 1994a, Scott 1998, and the methods mentioned above). However, the really difficult task is to label that relation between the terms. The importance of this step lies in major part because it acts both as a qualitative evaluation on the effectiveness of a method which merely associates two terms, and as a step towards a more fully specified taxonomy/ontology where the the nature of relations are explicit. Only if relations are explicit can an ontology be used with problem solving methods (PSMs) (Gomez-Perez 1999) i.e. for some form of logical inference.

A handful of efforts exist in the literature to identify terms in texts together with the specific relation. One approach is to focus on a specific semantic relation such as synonymy and this is the approach taken by Church *et al* (1994) although they prefer to use the term 'substitutability'. Another, well-known attempt, was that of Hearst (1992) who attempted to find lexical environments which would identify terms with the 'hyponymy' relation. She proposed five possible 'lexico-syntactic patterns' for identifying hyponyms. For example:

(1) *such NP as {NP, }*{(or|and)} NP*
… works by such authors as Herrick, Goldsmith and Shakespeare …

(2) *NP {,NP}*{,} or other NP*
…. Bruises, wounds, broken bones or other injuries ….

This work was taken up later on by Morin (1999a, 1999b) who proposed a development environment to sequentially identify a semantic relation, find the lecicosyntactic pattern for that relation and then search a corpus for corresponding text. He states that it "can find only a small portion of related terms due to the variety of sentence styles and the inability to find a common environment to all those sentences" (Finkelstein-Landau and Morin 1999:6).

6. Conclusions and Future Directions

Most existing methods are highly problematic for a variety of reasons. It is highly unlikely that a unique simple method can be found which will fulfil the criteria described above and also permit the labelling of relations. The approach we conclude that is necessary, and the one we will espouse in our future research is one which combines different methodologies in order to balance the weaknesses of one with the strengths of another.

Building on Hearst's and Morin's work, and recognising that Hearst's 'lexico-syntactic patterns' correspond to the templates so often used in Information Extraction (IE), we intend to use adaptive IE techniques (Ciravegna 2001) and Machine Learning to 'learn' the patterns which identify specific types of relations (such as hyponymy or meronymy). Such an approach, given a appropriate training corpus will be able to learn the patterns for a large variety of relations and types of terms. The 'lazy NLP' approach espoused by Ciravegna makes it possible to identify the most relevant level of data (word

[1] This is a corpus used in the AKT project consisting of the journal Nature from 1997-2001, 13000 files totaling approx. 20m words

form, punctuation, POS, parsed structure) which is appropriate for the system to learn from.
We also plan to use existing ontologies built by hand (such as the Gene Ontology) in order to provide both a gold standard for evaluative purposes, and also a data set of terms which are candidates for identifying the lexico-syntactic environments where their relations are apparent from a machine readable perspective.

Acknowledgements

Our thanks go to Prof. Yorick Wilks for advice and moral support, and to Dr. Fabio Ciravegna for his care and insight. This research is sponsored by the AKT (Advanced Knowledge Technology) project, funded by the EPSRC (GR/N15764).

References

Berners-Lee, T., J. Hendler, O. Lassila, (2001) *The Semantic Web*, in Scientific American Issue 501 (http://www.sciam.com/2001/0501issue/0501berners-lee.html)

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

Buzan, A., (1993) *The Mind Map Book* BBC Consumer Publishing: London

Church, Kenneth and W. Patrick Hanks, (1990), Word Association Norms, Mutual Information, and Lexicography, *Computational Linguistics*, 16, 22-29

Church, Kenneth, William Gale; Patrick Hanks; Donald Hindle and Rosamund Moon, (1994) *Lexical Substitutability*, in Atkins and Zampoli eds. "Computational Approaches to the Lexicon, Oxford University Press: Oxford pp.154-180

Ciravegna, Fabio, (2001), *Adaptive Information Extraction from Text by Rule Induction and Generalisation* in "Proceedings of 17th International Joint Conference on Artificial Intelligence (IJCAI 2001)" , Seattle, August 2001.

Cover, Thomas M., and Joy A. Thomas, (1991), *Elements of Information Theory*, John Wiley and Sons.

Cutting, Douglas R., David R. Karger, Jan O. Pedersen & John W. Tukey, (1992) *Scatter/Gather: A Cluster-based Approach to Browsing Large Document Collections*. In "Proceedings of the 15th International SIGIR '92", Denmark.

Davenport, Thomas H. (1998) *Some Principles of Knowledge Management* available at http://www.bus.utexas.edu/kman/kmprin.htm

Ericsson, K. A. & H. A. Simon, (1984) *Protocol Analysis: verbal reports as data.* MIT Press: Cambridge, Mass.

Fensel, D., F. van Harmelen, I. Horrocks, D.L. McGuiness, P.F. Patel-Schneider, (2001) *OIL: An Otology Infrastructure for the Semantic Web*, IEEE Intelligent Systems, 16: pp.38-45

Finkelstein-Landau, Michal, Emmanuel Morin, (1999) *Extracting Semantic Relationships between Terms: Supervised vs. Unsupervised Methods*, in "Proc. International Workshop on Ontological Engineering on the Global Information Infrastructure", 71-80, Dagstuhl Castle, Germany,

Gale, William, Kenneth Church & David Yarowsky, (1992) One Sense per Discourse, in *Proceedings of the ARPA Workshop on Speech and Natural Language Processing*, pp.233-237

Gilchrist, Alan and Peter Kibby, (2000), *Taxonomies for Business: access and connectivity in a wired world.* Report published by TPFL. http://www.tfpl.com/about_TFPL/reports___research/reports___research.html

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, 16-21 October 1999

Grefenstette, Gregory, (1994), *Explorations in Automatic Thesaurus Discovery*, Kluwer.

Gruber, T.R. (1993), *Towards Principles for the Design of Ontologies Used for Knowledge Sharing*. In Roberto Poli Nicola Guarino, ed., *Proc. of the International Workshop on Formal Ontology*, Padova, Italy

Guarino, N. and Welty, C. 2000. *Identity, Unity, and Individuality: Towards a Formal Toolkit for Ontological Analysis*. In H. Werner (ed.) ECAI-2000: The European Conference on Artificial Intelligence. IOS Press, Berlin, Germany: 219-223.

Hearst, M.A., (1992) *Automatic Acquisition of Hyponyms from Large Text Corpora*, in "Proc. of COLING 92", Nantes

Järvelin, K. & Kekäläinen, J. (2000). *IR evaluation methods for retrieving highly relevant documents*. In "Proceedings of 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval", Athens, Greece, 24.-28.7.2000, pp. 41-48.

Krovetz, Robert, (1998) *More than One Sense Per Discourse*, in "Proceedings of the ACL-SIGLEX Workshop", 1998

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

Morin, Emmanuel, (1999a), *Des Patrons lexico-syntaxiques pour aider au depouillement terminologique*, Traitment Automantique des Langues, 40:1, 143-166,

Morin, Emmanuel, (1999b), *Using Lexico-Syntactic patterns to Extract Semantic Relations between Terms from Technical Corpus*, in "Proc. of TKE 99", 268-278, Innsbruck, Austria,

Sanderson, Mark, and Bruce Croft, (1999) *Deriving concept hierarchies from text*, in "Proceedings of the 22nd ACM SIGIR Conference", pp. 206-213,

Schwaneveldt, R.W. (ed.),(1990) *Pathfinder Associative Networks: Studies in Knowledge Organization*, Intellect Books: Bristol, UK

Sinclair, J.M. (1987) *Looking Up*. Collins: London

Stutt, A., and E. Motta, (2000) *Knowledge modelling: an organic technology for the knowledge age*, in M. Eisenstadt, T. Vincent (eds.) "The Knowledge Web" Kogan Page: London