Ontologies, Taxonomies, Thesauri: Learning from Texts

Christopher Brewster and Yorick Wilks

Department of Computer Science,

University of Sheffield, Sheffield, UK

C.Brewster|Y.Wilks@dcs.shef.ac.uk

Abstract

The use of ontologies as representations of knowledge is widespread but their construction, until recently, has been entirely manual. We argue in this paper for the use of text corpora and automated natural language processing methods for the construction of ontologies. We delineate the challenges and present criteria for the selection of appropriate methods. We distinguish three major steps in ontology building: associating terms, constructing hierarchies and labelling relations. A number of methods are presented for these purposes ut we conclude that the issue of data-sparsity still is a major challenge. We argue for the use of resources external tot he domain specific corpus.

1 Introduction

For a period, during the 16th and 17th centuries, many thinkers and scientists, ranging from Bacon to Leibniz, were pre-occupied with the attempt to create "philosophical languages". These would provide the means to communicated perfectly philosophical and scientific ideas, and thus avoid the vagueness of human language. The greatest English exponent of this movement was Dr. John Wilkins, Bishop of Chester, founder of the Royal Society, who wrote a large volume entitled *An essay towards a real character and philosophical language*, published in 1668. Wilkins realised that, in order to create a perfect language, where each 'word' would refer to one unique item or idea, he would have to catalogue the whole of known human knowledge, and this is what he attempted by writing his book.

This work was later to inspire Roget in his famous *Thesaurus*, but also was a forefather of modern day effort at knowledge representation such as Cyc (Lenat et al. 1994) and Internet portals such as Yahoo[1] or the Open Directory[2]. From a slightly different perspective, the efforts at universal languages were forerunners of the great library classification systems such as Dewey and the Library of Congress.

As a founding member of the Royal Society, one of Wilkins' prime concerns was that the society should use his 'real character' and keep it up to date. But no one was willing to continue Wilkins' herculean task. Thus it is that in fields which attempt some form of representation of human knowledge, we encounter the double problem of knowledge acquisition and knowl-

[1] www.yahoo.com

[2] www.dmoz.org

edge maintenance. In contemporary times, we encounter these problems in Artificial Intelligence, Knowledge Management and the Semantic Web. The main focus of knowledge representation is currently in the construction and population of *ontologies*, which can be described as more formal versions of traditional taxonomies or Roget-type thesauri[3](Gruber 1993). The importance of the formal aspect lies in making them machine readable and thus allowing computers to perform operations over the information modelled in them. For the Semantic Web, ontologies are especially important as they enable the localised representation of world knowledge that personal agents (for example) need to operate on the Semantic Web, and they make possible a variety of reasoning services. They will provide the semantics for the annotations associated with each and every page of information on the Semantic Web. From a commercial perspective, 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 (Staab et al. 1999), 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 shown (Fensel et al. 2001).

Whatever the discipline, however, existing work on the construction of ontologies has concentrated on the formal properties and characteristics that

[3] This is in contrast to the 'controlled vocabulary' sense of thesaurus in library science.

an ontology should have in order to be useful (Gómez-Pérez 1999, Guarino 1998) rather than the practical aspects of constructing one. There is an assumption behind the Semantic Web that the knowledge needed for it to function will be obtained through manual labour on the part of users. It is assumed that people will individually annotate their own web pages with suitable 'semantic' tags which will enable machines to 'read' them. Equally, ontologies are assumed to be hand-crafted knowledge artifacts; the result of as much effort as a dictionary or encyclopedia.

This paper takes the approach that, given the 'info-smog' we live in (AKT 2001), hand-crafting is impractical and undesirable. While it is still a major research challenge to construct ontologies entirely automatically, the current tools available from the Natural Language Processing community make it possible to automate the task to a large extent and reduce manual input to where it makes the most qualitative difference. In Section 2, we describe discuss in greater detail the problem of manually constructing ontologies and argue for the use of text corpora as the main source of knowledge. In Section 3, we present a number of criteria as a guide for the method that need to be used for the automation of ontology construction. In Section 4, we present a number of methods for constructing ontologies from texts based on the criteria presented. Section 5 considers how to bridge the gap between the implicit knowledge assumed by a given text and the actual explicit knowledge present in the texts.

2 Problems with Knowledge Acquisition

Knowledge, it 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 (cf. Figure 1). 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 (Schvaneveldt 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. Gómez-Pérez (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

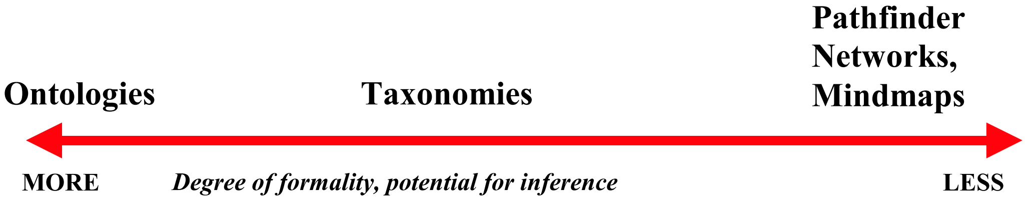
Ontologies
Taxonomies
Pathfinder Networks, Mindmaps
MORE
Degree of formality, potential for inference
LESS

Figure 1:

any experience lexicographers knows are just that: ‘ideals’.

One of the major problems in this field is that it is a common assumption among authors working with ontologies 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 & Motta 2000). 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 & 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[4], 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. 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 documents. 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 & Kibby 2000) . 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

[4] http://www.geneontology.org/

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[5], 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, 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,

[5] www.dmoz.org

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. 2001), and this we believe is the direction to take.

There are two traditional methods in artificial intelligence and knowledge management 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[6]. 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 lan-

[6] http://www-ksl-svc.stanford.edu:5915/

guage 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. 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 'markup' 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 a 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 (Gómez-Pérez 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: in a thesaurus such as Roget (Roget 1852) 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 (Fellbaum 1998), 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. But in a taxonomy or browsable hierarchy, we find concepts or terms are organised for the purpose of finding firstly relevant subcategories, and secondly specific web sites. Thus in Yahoo under Education → Higher Education → Universities we find a list of universities not a list of synonyms for the concept university. Linguistic coherence can be expected to be much more stable over time than encyclopaedic coherence, partly because language changes relatively slowly, and partly because our knowledge or understanding of the world tends to be revised rather dramatically in the light of social and cultural influences.

Wordnet has been criticised for the 'Tennis Problem' where terms associated with a particular topic are found in disparate places in the hierarchy of concepts. Thus 'tennis' is not near 'court' or 'tennis ball' or 'racket'. This criticism (Hayes 1999, Stevenson 2002) is the reflection of a particular set of expectations, a particular view point on how terms should be organised. It follows therefore that the notion of coherence must in effect be a specification of user requirements in terms of their unique perspective on the knowledge represented in the ontology.

Given these observations, the notion of coherence must be understood as being application specific. For our purposes in constructing taxonomies for Knowledge Management, 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 is 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 seemingly 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, or that 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/ Multiple Inheritance

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 some dictionaries by means of a subscript: class1, class2, class3, 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 (which determines 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. Similarly, pulmonary tuberculosis is both a respiratory disorder and an infectious disorder.

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. Noy & McGuiness (2001) term this phenomenon ‘multiple inheritance’ and give it a recognised place in their methodology. The problem really arises due to the bluntness of automated methods.

3.3 Ease of Computation

One of the major issues in Knowledge Management is the maintenance of the knowledge bases constructed. 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 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 & Croft (1999) discuss the difference and 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 Center 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 government which while comprehensible is not very easy to use and is discouraging for most users (Sanderson & Croft 1999). 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.

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 mentioned above. Secondly, it should permit the inclusion of an existing taxonomy (a 'seed') as a data structure to either revise or build upon as required

Ontologies and taxonomies are often legacy artifacts 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 world-view 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 From domain corpora to ontologies

In this section we present some of the key technologies to be used for the construction of ontologies in an automated or semi-automated manner. We take for granted here a step which is nonetheless important - term recognition. For each domain, for which we wish to build an ontology, the set of relevant terms needs to be identified because these will 'label' the concepts of importance in the domain, or be instantiations of particular concepts. Considerable research has been done in this field but we will not consider it here.

4.1 Associating Terms

The basic step in constructing an ontology is identifying which terms are associated with which. We need to know this in order to hypothesise likely candidates for the immediate ontological neighbours for each concept. There exist a large number of methods all of which take for granted the 'distributional hypothesis' which states that terms with a similar distribution (or behaviour) in texts are semantically similar or semantically related. Here we will just describe two to give a flavour.

Scott (1997, 1998) has shown that it is possible to derive a set of associated words for a given word using an extension of tf.idf measures. Thus, by comparing the frequency of a word in a given document with its frequency in a general reference corpus, it is possible to determine whether this is a key word or not, i.e. a word with a proportionately higher frequency. It is possible to construct thereby a set of key words for any given document.

N	WORD	No. of Files	AS
1	CHEESE	12	100.00
2	BRITANNICA	11	91.7
3	COM	10	83.33
4	MILK	5	41.67
5	WHEY	3	25.00
6	CHEESES	3	25.00
7	ACID	2	25.00
8	RIPENING	2	16.67
9	ROQUEFORT	2	16.67
10	CURD	2	16.67
11	EMMENTALER	2	16.67

Table 1: Associated Words for cheese using Encyclopaedia Brittanica texts

By extension, we can analyse all documents where a given word is a key word and identify the set of key words which are key in all these documents. These we call the key-key words. The associates of any given key-key word are those key words which co-occur in a number of texts. Example results are shown in Table 1 for the word *cheese* using texts from the Encyclopaedia Britannica.

Another classic approach is that proposed by Grefenstette (1994). This work rejects a variety of other computational methods such as those of (Brown et al. 1992). Grefenstette's approach is based on part-of-speech parsing and a similarity metric. The method that is employed by Grefenstette in his system is based on identifying the syntactic context for each word of interest, and then using each word in the context which bears a particular syntactic relation to the node as a semantic clue. Words which share such 'clues' are held to be alike. Node words are compared on the basis of the number of attributes ('clues') the two words share and the relative importance of the attributes. Example results are shown in Table 2.

word [Contexts]	Groups of most similar words
tissue [350]	cell | growth cancer liver tumor | resistance disease lens serum
treatment [341]	therapy | patient administration case response | result effect
concentration [339]	level content | excretion value | rate ratio metabolism synthesis
defect [338]	disturbance case malformation | regurgitation type response
rat [331]	animal mouse | dog mice level | infant kidney day rabbit group
method [298]	technique | procedure test | mean result study | group treatment

Table 2: From Grefenstette 1994: 51

It has to be stressed that all such methods cannot do anything more than provide candidates for the construction of the ontology. There is inevitably a lot of noise. For example, Grefenstette's approach places positive and negative adjectives together. More importantly, such methods, while suggesting that there *exists* a relation, do not tell *what* that relation is.

4.2 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. Considerable 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 & Hanks 1990, Cover & 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 & Smith (1996) take a closely related approach where the whole vocabulary is assigned to one class and this is divided into two subclasses which maximize 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 & 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"

Figure 2: A fragment of concept hierarchy

Figure 3: Another fragment of concept hierarchy

(This is not to be confused with the traditional notion of subsumption which refers to the ISA/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 frequent in most texts than dog. For example, in a corpus of texts from the journal

Nature, 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. 2, but Sanderson and Croft's approach would predict a structure as in Fig. 3. Thus it would appear that the Sanderson and Croft approach may be useful under certain circumstances but it does not provide clearly coherent output.

4.3 Labelling Relations and Finding Explicit Knowledge

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 as 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.

There are two approaches which one could take and these correspond to different strands in the relevant literature on the subject. One approach selects an ontological relationship (synonymy, or hyponymy, let us say) and attempts to develop an algorithm to identify terms whose relationship corresponds to the relation identified. Such an approach is taken by (Church et al. 1994) in searching for 'substitutability' which they note does not correspond directly with synonymy or any other classical semantic relationship. Having parsed a corpus, they analyse the objects of verbs and compare the overlap using the t-test. This enables them to provide a score of how appropriate it is to substitute one verb with another.

Hearst (1992) presents the most influential approach in this strand. She identified a number of lexico-syntactic patterns such as those shown in Table 3 which would allow one to identify pairs of terms standing in a specific ontological relationship. Although Hearst herself did not implement the idea, the work has underpinned much subsequent research on ontology building (Morin 1999, Brewster et al. 2002) and also ontology population (Cimiano et al. 2004).

Relation	Pattern (lexico-syntactic)	Example
HYPERONYMY	such NP as {NP, } * {(or|and)} NP NP , NP* , and other NP	such cars as the Mercedes C-Class, the Lexus ES 300 Ferrari, Honda, McLaren, Porsche, and other cars
MERONYMY	NP's NP, *	car's cooling system/ car's gas tank/ etc.

Table 3: Some Hearst patterns

The main problem with this approach is that of data sparsity in that it is quite hard to find both sufficient lexico-syntactic environments *per se* and sufficient exemplars or citations which include any two terms one is interested in. This problem is discussed further in Section 5.

The other approach, exemplified for example by the ASIUM system (Faure & Nédellec 1998) or CAIULA (Basili et al. 1996) has focussed on learning ontological information from the detailed syntactic analysis of corpora, especially verb subcategorisation frames. The output of such systems is very specific and it has yet to be shown by researchers that a subcategorisation based approach can lead to useful results for ontology building. Part of the problem lies in the fact that such approaches learn that term a and term b have a specific verb mediated relation or set of relations. This is specific to the context in which the terms were found and cannot easily generalise.

5 Data Sparsity and Dealing with Implicit Knowledge

In outlining the approach presented above, one major factor has been ignored which is that authors rarely make explicit ontological statements. We must remember that an ontology is a "shared conceptualisation" (Gruber 1993) i.e. consists of a shared set of concepts. For any given domain, the ontology is supposed to represent the concepts which are held in common by the participants in that domain. Thus it would appear that an ontology represents the background knowledge associated with a domain.

A domain specific text has a special relationship with that domain's ontology. When a writer creates a text they assume a number of things. There is a linguistic assumption concerning the language used and a cognitive assumption concerning the ability of the audience to understand the vocabulary and technical terms used. In effect, a writer assumes that the audience shares the same or almost the same knowledge as themselves.

This has an important consequence, generally ignored by researchers in automated ontology learning, which is that the background knowledge captured in an ontology is rarely explicitly stated in a text. It is implicit and taken for granted by the author. Consequently, it is very difficult to construct a computational process which will capture what in essence is not there. By explicit, we mean that a ontological relationship between two terms is expressed in some lexico-syntactic pattern of the type first identified by Hearst and discussed above.

Our research has shown that for domain specific corpora it is extremely improbably, irrespective of size, that sufficient exemplars will be found of any given lexico-syntactic pattern so as to be able to obtain reliable results (Brewster et al. 2003). Thus for a collection of texts from the journal *Nature* concerning genetics almost no exemplar contexts were found for randomly chosen pairs of terms from the *Gene Ontology*. The solution from an engineering point of view lies in finding alternative sources of ontologically explicit data. Such may be the Internet or specific sources relevant to the domain (possibly textbook, glossaries etc.). However, there is the need for a considerable amount of work before such a 'look elsewhere' approach can be made to work practically.

6 Conclusion

In this paper we have considered some of challenges concerning the construction and maintenance of taxonomies and ontologies, and we have argued for the use of texts as the basic resource for building them. We have proposed some criteria in the selection of appropriate methodologies for automating the ontology building task. We distinguished three major steps in the ontology building process and briefly considered some typical technologies which can be used for those purposes. We concluded by arguing that the fact that ontologies represent ‘background knowledge’ assumed by writers makes them unlikely to explicitly define their terms. This is a further engineering challenge to be overcome.

The major challenges have been outlined in this paper. Considerable effort is being expended by researchers to overcome these issues but their success can only be gauged if there are appropriate evaluation metrics such as those developed for classic Information Retrieval tasks (the TREC series, for example). Although some efforts are being made (Brewster et al. 2004), there is a great need for a standardisation process which will encourage measurable progress to be made.

Acknowledgements

This work has been carried on in the framework of the Advanced Knowledge Technologies (AKT) project (AKT 2001), 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.

References

AKT (2001), ‘Advanced knowledge technologies: Manifesto’, Available at www.aktors.org.

*http://www.aktors.org

Basili, R., Pazienza, M.-T. & Velardi, P. (1996), *A context driven conceptual clustering method for verb classification*, MIT Press, Cambridge, MA, pp. 117–142.

Berners-Lee, T., Hendler, J. & Lassila, O. (2001), 'The semantic web', *Scientific American* pp. 30–37.

Brewster, C., Alani, H., Dasmahapatra, S. & Wilks, Y. (2004), Data-driven ontology evaluation, *in* 'Proceedings of the 4th International Conference on Language Resources and Evaluation', European Language Resources Association, Lisbon.

Brewster, C., Ciravegna, F. & Wilks, Y. (2001), Knowledge acquisition for knowledge management: Position paper, *in* 'Proceeding of the IJCAI-2001 Workshop on Ontology Learning', IJCAI, Seattle, WA.

Brewster, C., Ciravegna, F. & Wilks, Y. (2002), User-centred ontology learning for knowledge management, *in* 'Proceedings of the 7th International Conference on Applications of Natural Language to Information Systems', Vol. 2553 of *Lecture Notes in Computer Sciences, Springer Verlag*, Springer-Verlag, Stockholm.

Brewster, C., Ciravegna, F. & Wilks, Y. (2003), Background and foreground knowledge in dynamic ontology construction, *in* 'Proceedings of the Semantic Web Workshop, Toronto, August 2003', SIGIR.

Brown, P., Pietra, V. D., deSouza, P., Lai, J. & Mercer, R. (1992), 'Class based n-gram models of natural language', *Computational Linguistics* **18**(4), 467–479.

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

Church, K. W. & Hanks, P. (1990), 'Word association norms, mutual information, and lexicography', *Computational Linguistics* **16**(1), 22–29.

Church, K. W., Hanks, P., Hindle, D., Gale, W. A. & Moon, R. (1994), Substitutability, *in* B. Atkins & A. Zampoli, eds, 'Computational Approaches to the Lexicon', OUP, pp. 153–80.

Cimiano, P., Handschuh, S. & Staab, S. (2004), Towards the self-annotating web, *in* 'Proceedings of the 13th International World Wide Web Conference, WWW 2004, New York, USA, May, 2004', ACM Press.

Cover, T. M. & Thomas, J. A. (1991), *Elements of Information Theory*, 2 edn, John Wiley.

Cutting, D. R., Karger, D. R., Pedersen, J. O. & Tukey, J. W. (1992), Scatter/gather: A cluster-based approach to browsing large document collections, *in* 'Proceedings of the Fifteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval', Interface Design and Display, pp. 318–329.
*http://www.acm.org/pubs/articles/proceedings/ir/133160/p318-cutting/p318-cutting.pdf

Davenport, T. H. (1998), 'Some principles of knowledge management'.
*http://www.bus.utexas.edu/kman/kmprin.htm

Dom, B. (1999), Automatically finding the best pages on the world wide web (clever), *in* 'Search Engines and Beyond: Developing efficient knowledge management systems', Boston, MA. April 19-20, 1999.

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

Faure, D. & Nédellec, C. (1998), A corpus-based conceptual clustering method for verb frames and ontology acquisition, *in* 'Proceedings of the LREC workshop on Adapting lexical and corpus ressources to sublanguages and applications', LREC, Granada, Spain.

Fellbaum, C. D. (1998), *WordNet: An Electronic Lexical Database*, MIT Press, Cambridge, MA.

Fensel, D., van Harmelen, F., Horrocks, I., McGuiness, D. L. & Patel-Schneider, P. F. (2001), 'Oil: An ontology infrastructure for the semantic web', *IEEE Intelligent Systems* **16**.

Gilchrist, A. & Kibby, P. (2000), *Taxonomies for Business: access and connectivity in a wired world*, TPFL Consultancy, London.
*http://www.tfpl.com/

Gómez-Pérez, 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'.

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

Gruber, T. R. (1993), 'A Translation Approach to Portable Ontology Specifications', *Knowledge Acquisition* **6**(2), 199–221.

Guarino, N. (1998), Some ontological principles for designing upper level lexical resources, *in* 'Proceedings of the First International Conference on Language Resources and Evaluation, LREC 98'.
*http://www.loa-cnr.it/Papers/LREC98.pdf

Guarino, N. & Welty, C. (2000), 'Identity, unity, and individuality: Towards a formal toolkit for ontological analysis', *Proceedings of ECAI-2000* . available from http://www.ladseb.pd.cnr.it/infor/ontology/Papers/OntologyPapers.html.

Hayes, B. (1999), 'The web of words', *American Scientist* **87**(2).

Hearst, M. (1992), Automatic acquisition of hyponyms from large text corpora, *in* 'Proceedings of the Fourteenth International Conference on Computational Linguistics (COLING 92), Nantes, France, July 1992'.

Järvelin, K. & Kekäläinen, J. (2000), 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', ACM SIGIR, ACM, Athens, Greece, pp. 41–48.

Lenat, D. B., Guha, R. V., Pittman, K., Pratt, D. & Shepherd, M. (1994), Cyc: Toward programs with common sense, Technical report, MCC and Stanford. *http://www.cyc.com/tech-reports/act-cyc-108-90/act-cyc-108-90.html

McMahon, J. G. & Smith, F. J. (1996), 'Improving statistical language model performance with automatically generated word hierarchies', *Computational Linguistics* **22**(2), 217–247.

Morin, E. (1999), 'Des patrons lexico-syntaxiques pour aider au depouillement terminologique', *Traitment Automantique des Langues* **40**(1), 143–166.

Noy, N. F. & McGuiness, D. L. (2001), Ontology development 101: A guide to creating your first ontology, Technical Report KSL-01-05, Stanford Knowledge Systems Laboratory, Stanford University, Stanford, CA.

Roget, P. M. (1852), *Thesaurus of English Words and Phrases Classified and Arranged so as to Facilitate the Expression of Ideas and Assist in Literary composition*, Longman, Brown, Green and Longmans, London.

Sanderson, M. & Croft, B. (1999), Deriving concept hierarchies from text, *in* 'Proceedings of the 22nd ACM SIGIR Conference', pp. 206–213.

Schvaneveldt, R. W. (1990), *Pathfinder Associative Networks: Studies in Knowledge Organization*, Ablex, Norwood, NJ.

Scott, M. (1997), 'PC analysis of key words - and key key words', *System* **25**(2), 233–45.

Scott, M. (1998), Focusing on the text and its key words, *in* 'TALC 98 Proceedings', Humantities Computing Unit, Oxford.

Sinclair, J., ed. (1987), *Look Up: An account of the COBUILD project in lexical computing*, Collins COBUILD, London.

Staab, S., Angele, J., Decker, S., Erdmann, M., Hotho, A., Mädche, A., Schnurr, H.-P., Studer, R. & Sure, Y. (1999), 'Semantic community web portals', *Computer Networks* **33**(1-6), 473–491. Special Issue: WWW9 - Proceedings of the 9th International World Wide Web Conference, Amsterdam, The Netherlands, May, 15-19, 2000.

Stevenson, M. (2002), Combining disambiguation techniques to enrich an ontology, *in* 'Proceedings of the Fifteenth European Conference on Artificial Intelligence (ECAI-02) workshop on "Machine Learning and Natural Language Processing for Ontology Engineering", Lyon, France'.

Stutt, A. & Motta, E. (2000), *Knowledge modelling: an organic technology for the knowledge age*, Kogan Paul, London, chapter 13, pp. 211–224.