Experiments in Ontology Construction from Specialist Texts

Mariam Tariq, Pensiri Manumaisupat, Rafif Al-Sayed, Khurshid Ahmad
Department of Computing
University of Surrey
{m.tariq, p.manumaisupat, r.sayed, k.ahmad}@surrey.ac.uk

Abstract

The identification of a domain ontology is usually a theoretical pursuit. However, the development of knowledge management systems and information extraction systems often requires an understanding of the ontology of the domain; the question of ontology then has a serious practical import. Knowledge is typically recorded in archives of texts: an examination of a sample of the archive may lead to the identification of a potential ontology of the domain. This is our claim.

1 Introduction

Text, whether in paper or electronic form, is a tangible source of information or knowledge that can be shared amongst a group of people in any domain. Almost all human enterprises are characterized by a text repository that is peer-reviewed and reflects the opinion of the majority of the enterprise; a repository may comprise books, learned papers, popular articles, dictionaries, manuals and handbooks. The advent of the World Wide Web has made it easier to collect, store and access such resources electronically.

The conceptual organization of any subject domain is based on a consensus amongst members of the domain. There are philosophical debates about how the concepts are organized and whether or not the organization is a social, psychological, or political phenomenon. The key for us is the existence of a consensus and the consensus manifests itself in the speech and writing of the domain community. The community, scientific, technical, recreational, political, religious or otherwise, continues to evolve and change. The evolution and change manifests itself in the linguistic output of the domain. This is not to deny that other semiotic systems, other than language, are not at work, but at least for us, linguistic output is amongst the most tangible.

The discussions of ontology, ontologies, and different things ontological, have a considerable psycho-philosophical overtone (Sowa 2000). This is perhaps due to the fact that unlike the building blocks of a subject domain – its *terms* – the *relationship* between the terms is not explicitly available. This is certainly true at the inception of a subject. When the subject matures the relationships are made more explicit and sometimes the conceptual organization is showed graphically as well. However, the relationship between terms is *signaled* in domain texts. There may be many complicated ways in which a natural language allows the description of the so-called semantic relationships; nevertheless it is not difficult to find rather straightforward ways used by the domain community to signal such relationships. For example, a forensic scientist describes a hyponymic relationship by the simple device of enumeration: *body fluids **like** saliva, blood, urine…..* and this practice of describing relationships through the use of phrases *like*, *including*, *such as, and/or other* is quite common in a number of typologically distinct domains. Cancer specialists write that *adjuvant therapies **include** chemotherapy and hormone therapy* and that *anthracyclines, docatexal **and other** drugs are used in chemotherapy.* Financial news wires contain examples of such signaling as in *defensive stocks **such as** oil, confectionary and electricity*. The semantic relationships are the basis

of the ontological commitment of a domain community. The inter-relationships of terms in a domain, when made explicit, for example, diagrammatically or more formally through a graph, will tell us something about the ontology of the domain.

In this paper we describe a method for identifying the ontological commitment of a domain by examining a random selection of documents produced by the members of the domain. The use of the various phrases to encode such complex relationships appears to have its own rules of description – a kind of *local grammar* governs the behaviour of clauses where the authors describe their ontological commitment. We describe how such structures have been identified in three different subject domains: *forensic science, breast cancer research* and *finance & business*, by exploring the local grammar. The penultimate use of making the ontological commitment explicit in the three domains is different.

2 Ontology Construction and Special Language Texts

We aim to explore the ontological commitment of a specialist domain from a randomly sampled domain-specific text corpus – that which might be construed to be a set of 'representative' texts of the domain. This notion of representativeness is controversial at best but has not deterred corpus linguists from building representative corpora of general language texts such as the British National Corpus (BNC) (Leech et al. 2001). Paradoxically this approach has been very productive; such corpora have led to new insights into the structure of language. The sampling is random in that the corpora are usually constructed through the use of domain-specific keywords in a Web search engine and texts of different genera are chosen. Specialist languages, considered variants of natural language, are restricted lexically, syntactically and semantically (Harris, 1988). Open class words dominate specialist language texts, particularly noun phrases (NPs); phrases used to name objects, events, actions and states relevant to the domain. It has been suggested that not only can terms be extracted from a specialist corpus (Ahmad & Rogers 2000, Bourigault et al. 2001) but also semantic relations of hyponymy and meronymy (part-whole relations) between terms (Ahmad et al. 2003, Hearst, 1992).

The open class words, particularly the single open class words, reflect the lexical choice of the domain measured by way of frequency of occurrence. Indeed, it has been claimed that the specialist texts contain the so-called *lexical signature* that distinguishes them from the everyday or general language texts. This signature can be elicited by comparing the frequency distribution of the open class words in a specialist corpus with that of the distribution of the same words in a 'representative' corpus of general language words, for example, the BNC. This measure has been referred to as the measure of *weirdness* (Ahmad & Rogers 2001) – the frequent use of such terms will appear unusual to a native speaker of English using the BNC as a standard. Terms with a high frequency and weirdness are usually considered good *candidate* terms. Domains are distinguished by the productive use of certain terms and, apart from inflectional and derivational use of these terms, much of the productivity manifests itself in the frequently used compound noun phrases that comprise one or more single-words that give the idiosyncratic lexical signature to a given specialist domain (see Table 1. below).

OCWs	WC	COMPOUNDS
Forensic Science Corpus (Total tokens: 610,197)		
evidence	21	crime scene
crime	53	forensic science
scene	38	law enforcement
forensic	471	crime scene investigator
analysis	11	workplace homicide
blood	12	supreme court
dna	33	crime scene photography
Breast Cancer Corpus (Total tokens: 226464)		
cancer	336	breast cancer
breast	749	breast cancer risk
women	16	metastic breast cancer
risk	47	breast carcinoma
patient	30	ovarian cancer
treatment	22	tamoxifen therapy
therapy	119	adjuvant therapy

OCW	WC	COMPOUNDS
Finance & Business Corpus (Total tokens: 681,215)		
percent	67	interest rate
million	13	million pounds
market	12	ftse index
pounds	32	percent rise
shares	104	tax profit
reuters	11222	recovery plan
company	5	central bank

Table 1. A comparison of high frequency OCWs (potential candidate terms) in three different domains showing the weirdness coefficient (WC) and frequent compounds

Once the single and compound words are identified automatically, by comparing the distributions in the specialist corpus with a representative general language corpus (for example forensic occurs 471 times more frequently in the forensic science corpus then it does in the BNC), one can then examine the ontological commitments by examining semantic relations amongst the terms. The terms in a domain are often related to each other through a range of semantic relations such as *hyponymy* and *meronomy*, which can be used to build hierarchies. These semantic relations are often exemplified in a language through the arrangement of certain terms in recurrent grammatical patterns that can be subsequently analyzed. Cruse (1986) has discussed the notion of *semantic frames*: a triplet of phrases - *X* REL *Y* where X and Y are noun phrases (NPs) and REL is a phrase generally expressed as *IS A, IS A TYPE OF/KIND OF* and *PART OF* for illustrating hyponymic and meronymic relationships respectively.

Apart from the signal cues mentioned above, which are used most commonly in biological classifications anyway, it has been suggested by Hearst (1992) that certain *enumerative* cues could also be used to identify such relationships through lexico-syntactic patterns occurring in texts such as the frame $(X_1 \ldots\ldots, X_n)$ *OR OTHER* Y where each X and Y are NPs and each X_i in the list $(X_1 \ldots\ldots, X_n)$ is a hyponym or subtype of Y. An example sentence here could be 'In the case of shootings or other fatal assaults' in which case a *shooting* is a type of *fatal assault*. The cue *and other* follows the same pattern as the *or other* cue whereas cues like *such as*, *including* and *like* have the order reversed where the NP representing the supertype occurs on the left with the list of subtype NPs on the right side of the cue, the last NP having an *and* or an *or* preceding it: Y *INCLUDING* $(X_1, \ldots\ldots \{or/and\}, X_n)$. An example sentence here could be "Trace evidence including fibers, hair, glass and DNA was found at the crime scene," *fibers, hair, glass* and *DNA* being types of *trace evidence*.

The phrases where the superordinate and subordinate/instances are linked together by one of the signals described above appear to be structurally similar to '*collocations [..] and frozen sentences [.....] one often encounters that cannot be related by formal rules of either phrase structure or transformational type*' (Gross 1993:26). For Gross, phrases used in telling time or idiomatic use of language require a finite state automata based on the so-called local grammar. Examples of the local grammar used for signalling relationships between one term (NP) and a set of other terms (NPs) are shown in Figure 1a and Figure 1b:

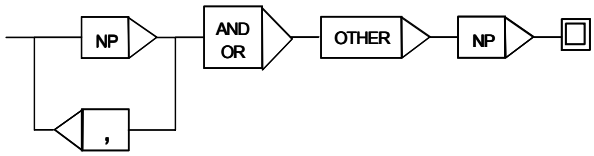

Figure 1a. Finite State Graph for the *and / or other* signal cues

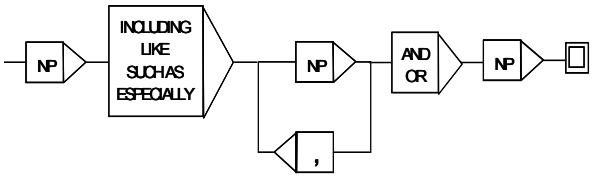

Figure 1b. Finite State Graph for the *including/like/such as/especially* signal cues

Compound words often convey a semantic relationship between the constituent lexical units as well. Compounding tends to *specialize* the meaning of the headword, each successive modifier specializing it further. This semantic relationship often signifies the hyponymic relation, for example, the compound term *trace evidence* suggests that *trace evidence* is a type of *evidence*. This heuristic can also be exploited to extract semantic relationships.

Sentences containing lexico-syntactic cues are automatically extracted and tagged to indicate the grammatical category of each word. Regular expressions are used to detect whether the local grammar is followed. All correct sentences are subsequently parsed, based on this local grammar, to extract hypernym-hyponym pairs. Compound NPs are also parsed recursively to extract more hypernym-hyponym pairs. All these hypernym-hyponym pairs extracted from the corpus are then merged together using a tree data structure that finally constitutes a forest of (sub-) trees, which can then be used in the identification of the ontological commitments of the domain. The final step is the representation of the various sub-trees in XML. An example of the process of analysing a sentence is shown in Figure 2.

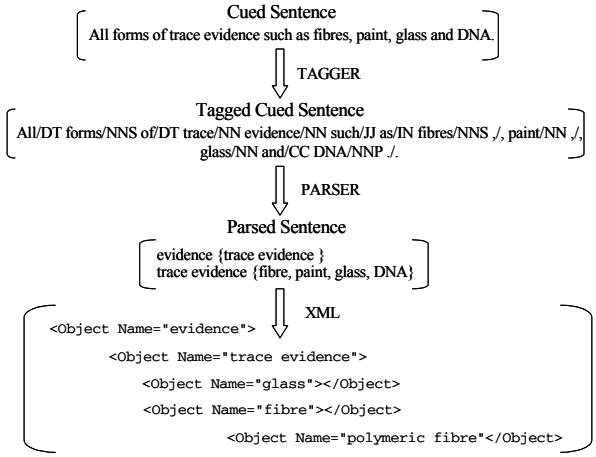

Figure 2. Analysis of a sentence containing a semantic frame to generate XML

Hence not only terms but also the relationships between the terms can be gleaned from domain specific texts. There are a number of other textual resources electronically available for certain specialist domains, including terminology databases or lexicons that can be used to validate the existence of the terms, and by the examination of term definitions provided, the relationships between different terms can be verified as well. This can be used as a part of an overall approach to extract terms and possible relationships. Furthermore there has been some extensive research done on creating general semantic lexicons like WordNet[1], as well as knowledge bases that claim to model world knowledge like CYC[2], which is still being developed and not freely available. Though useful for generic applications they are inadequate for use in specialized domains such as forensic science or medicine due to a lack in specialized terminology; knowledge will have to be acquired specifically for the specialized domain. As an example, highly salient terms in the forensic science domain such as *forensic science* and *crime scene* were not found in Wordnet. However, such knowledge sources may help in providing some top-level categories for a domain ontology. A system architecture for automatically extracting a domain ontology is shown in Figure 3.

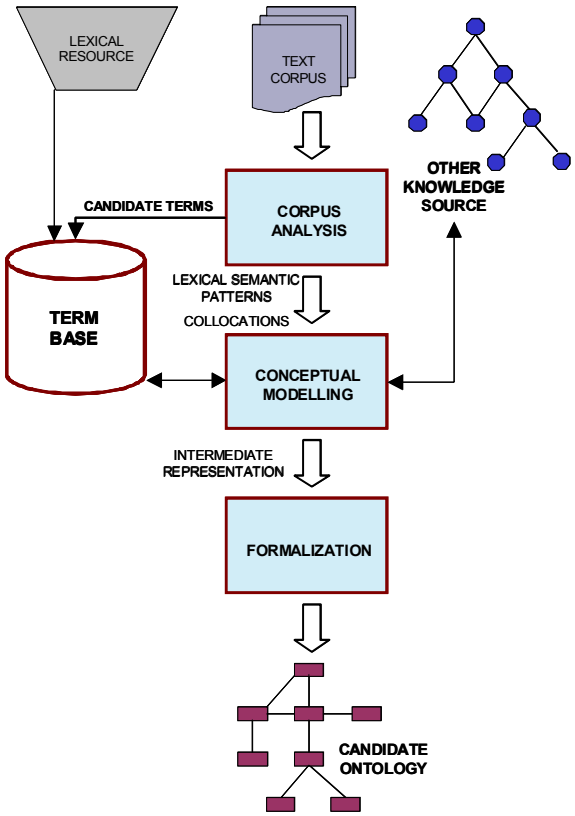

Figure 3. A proposed architecture for candidate ontology generation given a domain-specific text corpus

This architecture is based on a three-step method: (i) *knowledge acquisition* involves the

[1] http://www.cogsci.princeton.edu/~wn/
[2] http://www.cyc.com/

corpus creation and analysis phase; (ii) *conceptual modeling* involves the extraction of terms and their interrelationships as well as the integration of other knowledge sources and merging of partial knowledge structures into an intermediate representation; and finally (iii) *formalization* involves the mapping of the intermediate representation into a formalism such as XML. User interaction could be optional in each of the phases for validation purposes.

3 Randomly Sampled Domain-Specific Specialist Corpora and Ontology

In this section we shall provide some details of applying the method outlined in Section 2 on three domain-specific specialist corpora namely: *forensic science*, *breast cancer research*, and *finance and business*. We shall briefly describe the corpus and then some of the outputs from our method. In the *breast cancer* and *finance and business* domains there are online terminology databases available, which enabled us to compare our results to the term definitions provided for verification purposes.

3.1 Case Study 1: Forensic Science Domain

Content-Based Image Retrieval Systems are increasingly using keywords in addition to visual features for image categorization purposes (Squire et al., 2000). Texts related to images, also known as collateral texts, can help in the indexing and retrieval of images. Whereas *closely collateral* texts such as captions can be used to extract keywords to directly index or categorize images; we propose that *broadly collateral texts*, such as encyclopedic descriptions of objects within the image, may be a good source to extract related terms that can subsequently be used to build a thesaurus or ontology for query expansion purposes (Ahmad et al., 2003, Foskett 1997, Efthimiadis 1996). The method outlined in Section 2 was actually developed within context of the Scene of Crime Information System (SOCIS)[3] project, which attempted to exploit the use of texts related to images for the indexing and retrieval of crime scene images.

For the purpose of analyzing broadly collateral texts, a *forensic science* corpus of over half a million words was created from 1470 English texts (610,197 words). A variety of text types were collected from the Web such as journal papers, handbooks and advertisements ranging from 1990 to 2001 to ensure that the corpus was representative of the domain. Crime scene forms filled by Scene of Crime Officers were also included in the corpus.

The corpus had 20 OCWs amongst the first hundred with *evidence* and *crime* being most frequent. These terms are used productively to make compounds, for example *crime* and *scene* are used together to form 90 different compounds including *crime scene analysis* and *crime scene photography*. Some interesting candidate neologisms, which have a weirdness of infinity since they are not found in the BNC, included: *bitemark*, *toolmark*, *handgun*, *polygraph, footprints* (example of an unusual inflection) and *rifling* (example of an unusual derivation).

The forensic science corpus was then analysed to extract semantic relationships between the terms. Over 1200 sentences containing enumerative cues were extracted and it was observed that 60% of them incorporated the local grammar representative of hyponymic relationships between the phrases $(X_1,\ldots\{or/and\},X_n)$ and Y. It was interesting to note that the more typical X *IS A* Y frame pattern only brought up 40 valid sentences out of 400. Frames depictive of the meronymic relationship were not very productive; only 60 sentences were extracted with 40% of them being representative.

The diagram (Figure 4) shows a graphical view of some of the automatically extracted relationships. A partial hierarchy is shown of the concept *evidence* that has *trace evidence* as a sub-concept. *Blood*, *fibre* and *DNA* are types of *trace evidence* and *fibre* can be further classified as *inorganic, dye* or *manufactured polymeric fibre*.

[3] http://www.computing.surrey.ac.uk/ai/socis/

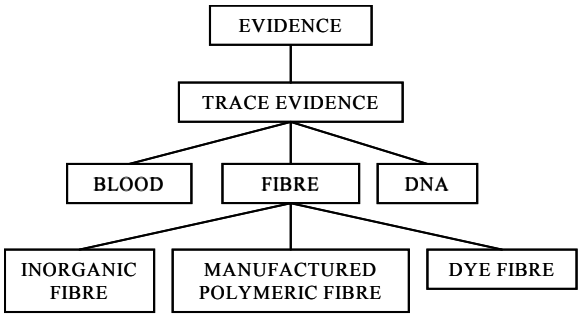

Figure 4. A partial-hierarchy of the concept *evidence* in the forensic science domain, automatically extracted from a randomly-selected corpus of texts

3.2 Case Study 2: Knowledge Maps for Breast Cancer Research

One can broadly divide the knowledge of an organization, or more specifically that of the people within it, into knowledge that is a result of *innovation*, either accidental or planned, and knowledge that comes about through the *application* of knowledge, the so-called *best practice*. The innovation is usually attributed to the R&D professionals (Huseman and Goodman, 1998) and *best practice* to the knowledge workers –professionals involved in running the day-to-day operations of complex organisations (Davenport and Laurence, 2000).

More often than not clarity in language-based communication is cited as the principal impediment in the exchange of knowledge, both innovative and extant, the later relates to best practice. There are best practices that have evolved in the care and treatment of virulent diseases like cancer. Early identification of symptoms by the potential patient is regarded as key to successful therapy; early notification of adverse effects, which could easily be reversed by changing the drug regimen, is crucial. In both these cases the knowledge of innovation and that of best practice has to be communicated across different registers: from the innovators to the professionals and practitioners and then onto the public at large.

One potential answer is to allow access to key documents that lead on from invention onto best practice; a document repository which can be searched through the use of terms in different registers and which is cross-referenced in a dynamic manner. Most importantly, the ontological basis of the repository should be open to inspection – *what* is included in the repository and *why*. One solution is the development of computer-based methods that can automatically index and cross-reference documents that make up the knowledge of a domain. We describe how knowledge within a given domain can be *mapped* so as it is available almost independent of the register. A *knowledge map*, a representation of concepts and their relationships, enables a user to navigate through the network and follow links to relevant knowledge sources (information or people) from any specific concept (Chou and Lin, 1998, Browne et al., 1997).

Our case study aims to describe a method for mapping knowledge in the sub-domain of breast cancer addressed specifically at the health professional level. The first step was to create a representative text corpus of nearly a quarter million tokens, from a variety of 1000 English texts (226,464 words) collected from the Web; mainly professional abstracts and journal papers from the NCI[4] website. An analysis of the corpus to determine the most frequent single and compound words can lead to the identification of a lexical signature, which can be used as a basis to develop the knowledge map of the domain. Amongst the 10 most frequent terms, *chemotherapy* (9th most frequent) appears nearly 756 times more frequently in the breast cancer corpus then it does in the BNC (recall Table 1, Section 2). Certain terms such as *estrogen* have a weirdness value of infinity, indicating that they may be candidate neologisms.

We have argued above that the constituent lexical units of a compound word or collocation frequently reflect a semantic relationship. For example, taking the compound term *breast cancer*, which has a high frequency of 1581 in the corpus, it can be deduced that *breast cancer* is a type of cancer, distinguishing it from say, *ovarian cancer*. This heuristic was used to indicate relationships between terms (Figure 5).

This method will also be applied to map knowledge from corpora comprising texts addressed at the patient and practitioner level respectively. A comparison between these different levels of knowledge will perhaps lead to

[4] http://www.cancer.gov/

a *common* knowledge map for breast cancer. Such a map could be used for generating text automatically for multi-level knowledge workers (health professionals, nurses and patients). We feel one of the main advantages of this method is its ability to be fully automated, using a database system as a backend. However, further evaluation and experimentation will determine its full potential for use in extracting domain knowledge for different levels of audience.

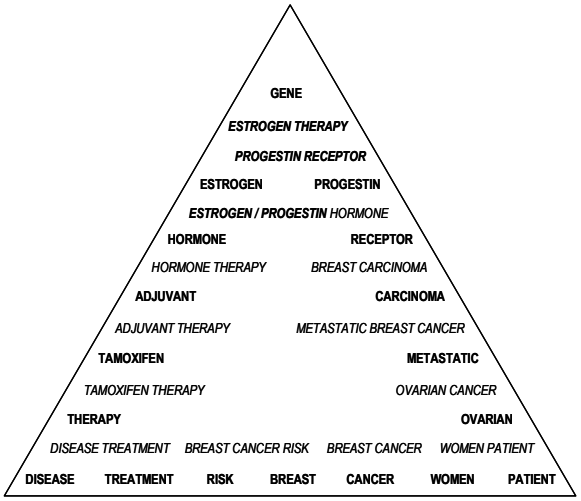

Figure 5. A Knowledge Map of some of the frequent NPs in the breast cancer domain shown in a triangle with the most frequent single-word terms at the base and higher weirdness towards the apex

The method outlined in Section 2 was applied to the breast cancer corpus, which can be considered highly specialized. Results from our analysis include (shown in XML):

```
<Object Name="adjuvant treatment">
    <Object Name="radiation" />
    <Object Name="chemotherapy">
        <Object Name="docetaxel" />
        <Object Name="anthracyclines" />
        <Object Name="vinorelbine" />
    </Object>
    <Object Name="hormone therapy" />
</Object>
<Object Name="local therapy">
    <Object Name="radiotherapy" />
</Object>
```

Note that the above relationship in fact elaborates further on the definition of *adjuvant treatment* that is given in the NCI terminology database. The definition in the database is a rather a limited one:

> **adjuvant therapy:** Treatment given after the primary treatment to increase the chances of a cure. Adjuvant therapy may include chemotherapy, radiation therapy, hormone therapy, or biological therapy

The following diagram (Figure 6) shows a partial-hierarchy automatically extracted from the corpus using both the local grammar and the compound term analysis. For example, from the following XML output it can be elicited that "atherosclerotic disease" is a type of disease and "breast cancer" is also a type of disease. In all approximately 70% of the extracted relationships were valid.

```
<Object Name= "atherosclerotic disease">
    <Object Name= "stroke" />
    <Object Name= "chd" />
</Object>
<Object Name= "disease">
    <Object Name= "breast cancer" />
</Object>
```

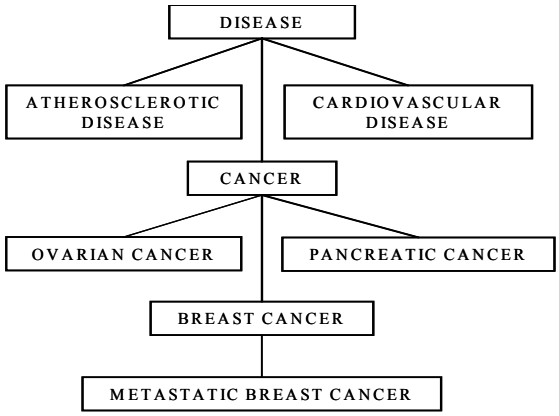

Figure 6: A partial hierarchy showing disease and some of its sub-concepts automatically extracted from the *breast cancer* corpus

3.3 Case Study 3: Finance & Business Domain

The construction of a *terminology database* for a given enterprise requires the identification of a *conceptual model* of the enterprise. The term conceptual model emphasizes that the relationships and interdependencies between the key ob-

jects in the enterprise be clearly delineated. Most conceptual models of terminology databases typically address this question at the level of *linguistic description*. The discussion here is on the attributes of the various linguistic 'slots' including lemma, definition, related terms, grammar categories/codes, as well as on administrative 'slots' such as dates of entry and update, and the names of the terminologists (Kugler et al., 1995).

The conceptual models for terminology databases seldom explicitly include a conceptual model of the domain in which the term base will be used. There is considerable discussion on the *concept-based approach* to terminology studies. This Platonist, neo-positive approach to knowledge was pioneered by Eugene Wüster and is continued to date in terminology literature (Cabre, 1999) and in terminology standards (ISO 1087-1/2:2000, ISO 12616:2002). The conceptual model is usually produced by a group of experts, typically working for a standards or trade organization, for an established or mature subject. The model is based on a consensus achieved over a period of time. Such considerations do not apply in a straightforward manner to an emergent subject domain. Here the experts are few and far between and as the subject is emerging there is inevitably not as much consensus as one may have for a maturer subject. However, experts still communicate through the medium of writing.

There are instances where the development of the subject is of direct interest to the public and newspapers and magazines tend to include the writings of the workers in an emergent domain. Financial market trading is a good example of an emergent subject where the subject involves academic and professionals publishing in journals, magazines and financial newspapers. Newer financial instruments – *shares*, *currencies*, *bonds,* good examples of a financial instrument – are being devised by financial traders while financial news reporters write about the state of these instruments. We describe how our 3-phase method of extracting terms and ontology structure may help in this emergent domain. The candidate ontology may be used as a basis of developing a fully-fledged conceptual model. The corpus was created by examining 1529 English texts (681,215 words) produced by Reuters[5] on the topic of British financial trading.

The analysis phase showed that of the 100 most frequent tokens in the corpus 42 were open class words including *percent, market, shares, company, bank, stock*, and *ftse*. Out of these 42, 60% were found as single terms or part of a compound term or phrase in the terminology base[6]. Most frequent compound terms extracted include *million pounds*, *ftse index*, *wall street*, *percent rise, tax profits* and *interest rates*.

The term *stock* is an important term in that it appears amongst the 100 most frequent terms in our corpus of British financial trading. One of the interesting collocates of the term was *defensive stock*. This is defined in the terminology dictionary as:

> '*a stock that tends to remain stable under difficult economic conditions. Defensive stocks include food, tobacco, oil, and utilities*. […]'.

Our ontological analysis throws interesting light on that. First, our analysis extends the definition to cover other than the four industries listed in the definition, and also the sub-specialisms of defensive stocks or sectors (cf: *confectionary* > {*chocolate, toffee*})

```
<Object Name="defensive stocks">
        <Object Name="electricity" />
        <Object Name="utilities" />
        <Object Name="oils" />
        <Object Name="soft drinks company" />
        <Object Name="tobacco" />
        <Object Name="water company" />
        <Object Name="banks" />
        <Object Name="utility" />
        <Object Name="retailers" />
        <Object Name="confectionery">
                <Object Name="chocolate" />
                <Object Name="toffees" />
        </Object>
</Object>
```

Second, our system can find instances of the type *defensive stock* as in the pharmacuetical company *Astrazeneca*.

[5] http://www.reuters.com/
[6] http://www.investorwords.com/

```
<Object Name="defensive-type stocks">
    <Object Name="astrazeneca" />
</Object>
```

Third, the opposite of the defensive stock or sector is the so-called *cyclical sector* defined as:

> '*The stock of a company which is sensitive to business cycles and whose performance is strongly tied to the overall economy. […] gains by buying the stock at the bottom of a business cycle, just before a turnaround begins.* ***opposite of*** *defensive stock*'.

Our system found one type of cyclical company (*mining*) as well as two instances of such stocks were found:

```
<Object Name="industrial cyclical companies">
    <Object Name="mining" />
</Object>
<Object Name="industrial cyclical stock">
        <Object Name="ici" />
        <Object Name="invensys" />
</Object>
```

There are few systems that produce a basis for generating conceptual models and our results encourage us to believe that our method will help here. The diagram below shows a partial hierarchy of the key term *stock*, the different types of stocks: *defensive* and *cyclical* as well as two instances of *industrial cyclical stocks*.

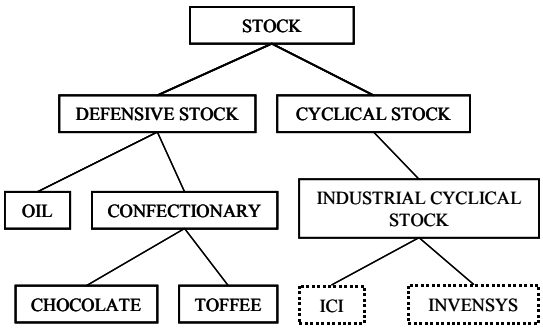

Figure 7. A partial hierarchy automatically extracted from the corpus of *finance and business*, showing some sub-concepts and instances of *stock*

4 Afterword

The three case studies discussed above illustrate that the method outlined for ontology construction showed promising results in three very disparate specialisms. This is a good indication that the method may be applicable to any *arbitrary* specialist domain. The resulting ontology could be subsequently used for a variety of purposes: query expansion, term base construction or knowledge mapping as discussed in our case studies.

If certain electronic lexical resources or term bases already exist for a certain domain then they can also be utilized to provide some additional information such as synonyms and alternate terms as well as validate the candidate terms and relationships that have been automatically extracted. In return our method can be used to periodically update the term base. The maintenance of term bases or domain ontologies if done manually is a difficult and time-consuming task. Since the introduction of neologisms as well as the obsolescence of certain terms is a common occurrence in many research active domains, there is a need for this problem to be addressed. The method for ontology construction could help to update a term base or ontology by analyzing current texts and adding any new terms found frequently and depopulating a term base or ontology of terms that have not been used for a certain period of time.

Taking into consideration Guarino's (1998) suggestion on developing different levels of ontologies depending on their generality, we can suggest that existing general lexicons such as WordNet or knowledge bases such as CYC could be used to provide a *top-level ontology* which our automatically constructed *domain ontology* could be merged with. The analysis of current texts such as the financial news texts, which can help provide current and often ephemeral relationships as well as instances of concepts can be used to build an *application ontology*, which might need to be updated frequently to reflect the changes in the state of the domain.

Acknowledgements

This work is related to two projects: SoCIS (GR/M89041), a three-year EPSRC sponsored project jointly undertaken by the Universities of Sheffield and Surrey and supported by five police forces in the UK; and GIDA (IST 2000-31123) a two-year EU sponsored project undertaken by University of Surrey in collaboration with EU companies. Mariam Tariq gratefully acknowledges a student bursary provided by the EPSRC.

References

Khurshid Ahmad and Margaret Rogers. 2001. *Corpus-based terminology extraction*. In: Budin, G., Wright S.A. (eds.): Handbook of Terminology Management, Vol.2. John Benjamins Publishers, Amsterdam. 725-760

Khurshid Ahmad, Mariam Tariq, Bogdan Vrusias and Chris Handy. 2003. *Corpus-Based Thesaurus Construction for Image Retrieval in Specialist Domains*, In (ed). Fabrizio Sebastiani. Proc. of ECIR'03. LNCS-2633. Springer Verlag, Heidelberg. 502-510.

D. Bourigault, C. Jacquemin, M-C. L'Homme, (eds.): 2001. Recent Advances in Computational Terminology. John Benjamins Publishers, Amsterdam.

G. Browne, S. Curley and P. Benson. 1997. *Evoking Information in Probability Assessment: Knowledge Maps and Reasoning-Based Directed Questions.* Managements Science (43:1). 1-14.

M. T Cabre. 1999. *Terminology: Theory, Methods & Applications.* Benjamins, John Publishing Company, Amsterdam. (Tr. Janet Ann DeCesaris).

Chou and H. Lin. 1998. *The Effects of Navigation Map Types and Cognitive Styles on Learners Performance in Computer-Networked Hypertext Learning System.* Journal of Educational Multimedia and Hypermedia (7). 151-176.

D. A. Cruse. 1986. *Lexical Semantics*. Cambridge University Press, Avon, Great Britain.

Thomas H. Davenport and Prusak Laurence. 2000. *Working Knowledge: How Organizations Manage What They Know.* Boston: Harvard Business School Press.

E. N. Efthimiadis. 1996. *Query Expansion.* In: M. E. Williams (ed.). Annual Review of Information Systems and Technology (ARIST). Vol.31. 121-187.

D. J. Foskett. 1997. *Thesaurus*. In: Sparck Jones, K., Willet, P. (eds.): Readings in Information Retrieval. Morgan Kaufmann Publishers, San Francisco, California. 111-134

Maurice Gross. 1993. *Local grammars and their representation by finite automata*. In: Hoey, M. P. (ed.): Data, Description, Discourse. HarperCollins, London 26-38.

Nicola Guarino. 1998. *Formal Ontology and Information Systems*. In Proceedings of FOIS'98 –formal Ontology and Information Systems. Trento, Italy, 6-8 June. IOS Press.

Z.S. Harris. 1988. *Language and Information*. In: Nevin, B. (ed.): Computational Linguistics Vol. 14, No.4. Columbia University Press, New York. 87-90

Marti Hearst. 1992. *Automatic Acquisition of Hyponyms from Large Text Corpora*. In Proceedings of the Fourteenth International Conference on Computational Linguistics (COLING'92). Nantes, France. 539-545.

Richard C. Huseman and Jon P. Goodman. 1998. *Leading with Knowledge: The Nature of Competition in the 21st Century*. Thousand Oaks: SAGE Publications.

ISO 1087-1:2000. Terminology work -- Vocabulary -- Part 1: Theory and application.

ISO 1087-2:2000. Terminology work -- Vocabulary -- Part 2: Computer applications.

ISO 12616:2002. Translation-oriented terminography

M. Kugler. K. Ahmad and G. Thurmair (Eds.) 1995. Translators Workbench: Tools and Terminology for Translation and Text Processing. Springer Verlag.

G. Leech, P. Rayson, A. Wilson. 2001. *Word Frequencies in Written and Spoken English: based on the British National Corpus*. Pearson Education Limited, Great Britain

John Sowa. 2000. *Knowledge Representation: Logical, Philosophical, and Computational Foundations.* Pacific Grove CA: Brooks/Cole.

McG. D. Squire, W. Muller, H. Muller, T. Pun. 2000. *Content-Based Query of Image databases: Inspirations from Text Retrieval*. Pattern Recognition Letters 21. Elsevier Science B.V. 1193-1198