Creating and Using Domain-Specific Ontologies for Terminological Applications

Diana Maynard¹, Sophia Ananiadou²

¹Dept. of Computer Science, Sheffield University Regent Court, 211 Portobello Rd Sheffield S1 4DP, U.K. D.Maynard@dcs.shef.ac.uk

²Computer Science, School of Sciences, University of Salford, Newton Building, Salford, M5 4WT, U.K. S.Ananiadou@salford.ac.uk

Abstract

Huge volumes of scientific databases and text collections are constantly becoming available, but their usefulness is at present hampered by their lack of uniformity and structure. There is therefore an overwhelming need for tools to facilitate the processing and discovery of technical terminology, in order to make processing of these resources more efficient. Both NLP and statistical techniques can provide such tools, but they would benefit greatly from the availability of suitable lexical resources. While information resources do exist in some areas of terminology, these are not designed for linguistic use. In this paper, we investigate how one such resource, the UMLS, is used for terminological acquisition in the TRUCKS system, and how other domain-specific resources might be adapted or created for terminological applications.

1. Introduction

Although in the past terminological applications such as automatic term recognition have been largely statistical, hybrid approaches combining linguistic and contextual information are becoming increasingly popular. One of the major obstacles facing the development of such systems is the absence of suitable high-quality specialised resources such as dictionaries, thesauri and ontologies. It is possible to acquire semantic information using purely corpus-based approaches (Soderland et al., 1995; Grefenstette, 1994), but there are many disadvantages, such as unreliable corpora and the difficulty in extracting the necessary information. In particular, where small corpora are concerned, statistical methods are unreliable, due to insufficient data.

Dictionary-based methods are widely used for general language applications, such as word sense disambiguation (Smeaton and Quigley, 1996; Yarowsky, 1992; Wilks and Stevenson, 1998), but are not so appropriate for terminological applications. There are two main reasons for this. Firstly, terms and words have different linguistic properties: for example, most terms consist of more than one word. Secondly, general dictionaries are not specific enough to deal with specialised terminology, and tend to have serious gaps in their coverage. Terminological applications almost always require that their lexical resources be tailored to the domain, but there is a lack of suitable specialised resources available. Even in the field of medicine, which is relatively rich in lexical resources, these may still be inadequate, because of the disparate nature of material covered in the field. The medical field is both multi-faceted and dynamic (Sager, 1990), and is therefore particularly susceptible to ambiguity. The necessary criteria imposed on lexical resources for such fields are also far more rigorous, ideally being organised for "optimum reference utility, readability, interchangeability and flexibility" (Lynch, 1997).

2. The UMLS Knowledge Sources

In the TRUCKS system (Maynard and Ananiadou, 1999a), the UMLS Knowledge Sources (NLM, 1997) are used to provide semantic information for term recognition and disambiguation¹. These contain information about medical terminology, organised in a hierarchical structure, which provides both a classification system for the terms, and relational information between them. It consists of four main components:

¹Version 4 was used for TRUCKS, although later versions have since been produced

- Metathesaurus
- Semantic Network
- Source Vocabularies
- Specialist Lexicon

The **Metathesaurus** contains semantic information about the terms that appear in various controlled vocabularies and classifications, such as SNOMED and MESH. It lists over 330,000 concepts and over 735,000 terms, from over 30 sources. The meanings and relationships are preserved from the source vocabularies, but some additional information is provided, and new relationships between concepts and terms from different sources are established.

The Semantic Network contains information about the set of basic semantic categories that may be assigned to the concepts, and about the relationships between these categories. In Version 4, there are 135 semantic types (represented as nodes) and 51 relations (represented as links). The main relation is that of hyponymy, but there are also 5 main types of non-hierarchical relations: physical, spatial, temporal, functional and conceptual. Relational information is generic rather than specific, i.e. details are provided about semantic types rather than about individual concepts, and relations do not necessarily hold for every concept belonging to that type. For example, a relationship exists between the semantic classes of disease or syndrome and acquired abnormality, such that the former is a *result_of* the latter. This does not imply that any disease is the result of any acquired abnormality, but simply that there is a general relationship of this kind between the two classes, such that diseases or syndromes can be the result of acquired abnormalities.

The other two sources contain detailed morphological and syntactic information about words and terms, and details of the original sources of all terms. These are not used in the TRUCKS system.

2.1 Semantic Information in the UMLS

The UMLS Semantic Network differs from traditional thesauri in two main ways. Firstly, it is non-lexicalised, providing a hierarchical structure of concepts (or semantic types), rather than terms. Files are provided in the Metathesaurus linking these concepts to the terms they represent. Secondly, the entries in the UMLS are derived from a number of different source vocabularies. The resulting ontology is thus a conglomeration of meanings, from a variety of viewpoints. The implication is that the concepts represented can only really be interpreted in their extensional meaning (whereas the source vocabularies from which they are derived provide the intensional meaning of the concepts (Campbell et al., 1998)). Campbell et al. propose that the concept in UMLS holds a special relationship with the meaning it represents, in that it takes on the "emergent" meaning, which may not be the same as the original meaning in the source vocabularies. In this way, the UMLS Semantic Network can be seen as a *possible world*.

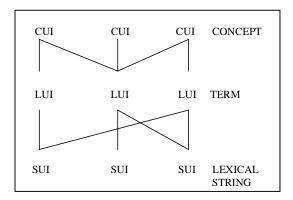


Figure 1: Relations between strings, terms and concepts

The UMLS makes a distinction between concepts (meanings) and concept names (terms). The Metathesaurus is organised by concept, establishing relationships between alternate names and views of the same concept, and between concepts themselves. The representation takes the form of three levels: a set of general concepts, represented by codes known as CUIs (Concept Unique Identifiers); a set of concept names, represented by codes known as LUIs (Lexical Unique Identifiers); and a set of strings, represented by codes known as SUIs (String Unique Identifiers). Figure 1 depicts the structure of the UMLS as a whole, showing how the lexical string is linked to the terms, which are then linked to the concepts, and how ambiguity and variation can occur. Term variation occurs where a concept is linked to more than one term. Term ambiguity occurs where a term is linked to more than one concept. Similarly, lexical variation occurs where a term is linked to more than one lexucal string, and lexical ambiguity occurs where a lexical string is linkes to more than one term. Table 1, taken from the UMLS documentation, shows a more concrete example of this. Atrial fibrillation and atrial fibrillations are different strings, but linked to the same term (lexical ambiguity). Atrial fibrillation and auricular fibrillation are different terms, but linked to the same concept (term ambiguity).

The way in which terms and concepts are represented in the UMLS makes it particularly interesting to use for term disambiguation, because it is non-lexicalised. The meaning of the terms lies in the source vocabularies from which the Network is compiled, rather than directly in the concepts represented by the terms, and therefore univocal correspondences between term and concept are not forged in the same way as in other thesauri and ontologies.

Concept (CUI)	Term (LUI)	String (SUI)
C0004238	L0004238	S001668
(preferred)	(preferred)	(preferred)
Atrial	Atrial	Atrial
fibrillation	Fibrillation	fibrillation
Atrial	Atrial	
fibrillations	fibrillations	
Auricular		S0016669
fibrillation		(plural variant)
Auricular		Atrial
fibrillations		fibrillations
	L0004327	S0016889
	(synonym)	(preferred)
	Auricular	Auricular
	fibrillation	fibrillation
		S0016900
		(plural variant)
		Auricular
		fibrillations

Table 1: Relationship between concepts, terms and strings

2.2 Using the UMLS for Similarity Calculation

The UMLS is chosen over other resources for TRUCKS for four main reaons.

- 1. The relationship it expresses between terms and concepts allows us to exploit the multidimensional nature of terms.
- 2. It is structured in such a way that it is very easy to manipulate.
- 3. It is the most comprehensive of the medical resources considered, since it not only contains all the SNOMED terms, but also information from many other sources.
- 4. It is easy to obtain for research purposes, and updates and backup support are available.

The UMLS Metathesaurus and Semantic Network are used in TRUCKS to allocate semantic categories to candidate terms, and to calculate similarity between term and context (Maynard and Ananiadou, 1999b, Maynard and Ananiadou 1999c). The similarity is measured between a context term and a candidate term by calculating the distance between their semantic categories in the Semantic Network. The semantic distance is defined in terms of two weights:

• **positional**: the vertical position of the Most Specific Common Abstraction of the two nodes², measured by the combined distance from root to each node

• **commonality**: measured by the number of shared common ancestors multiplied by the number of nodes being compared (usually 2).

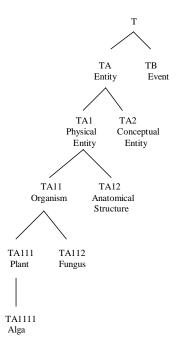
Figure 2 depicts a section of the UMLS Semantic Network, from which the similarity is calculated. Similarity between two nodes is calculated according to the following equation, to produce a score between 0 and 1:

sim(w1, w2) = com(w1, w2) / pos(w1, w2)

where:

com(*w1*, *w2*) is the commonality weight of word 1 and word 2

pos(w1, w2) is the positional weight of word 1 and word 2.





2.3 Shortcomings of the UMLS

Although the UMLS provides a simple, pre-defined hierarchy for similarity calculation, it does suffer from several drawbacks. Firstly, the design of any manually created hierarchy is subjective, and prone to errors and omissions. Secondly, it is not tailored to the corpus. Our results revealed many candidate terms selected by the SNC-Value method, and validated by human experts, which were not contained in the UMLS. This is due to a number of reasons, but largely because the UMLS is not specific to eye pathology (the domain in which we tested it) but only to medicine in general, and therefore terms

² (Kolodner, 1993) the lowest node which is an ancestor of both nodes

which are very specific to eye pathology are not always present. Also, there are some strange anomalies in the UMLS, possibly caused by the mapping of duplicate concepts in the source to single concepts in the UMLS. For example, the term *Esophageal cancer* (C0546837) appears in the relationship 'is a child of' to the concept *Esophageal cancer* (C0546837), where, by definition of their coding, the two concepts are clearly identical³.

One solution to these problems is to generate a thesaurus automatically from the corpus. For example, Ushioda's statistical algorithm for the hierarchical clustering of words (Ushioda, 1996) has been implemented in preliminary experiments to create a hierarchy for terms (Mima et al., 1999). Initial experiments have also been carried out using this method on the eye pathology corpus used for TRUCKS, to compare the ontology created with the UMLS.

An alternative is to use a pre-existing thesaurus, but to augment it with supplementary information from the corpus, using statistical methods. This has the advantage of making use of the accuracy and efficiency of predefined resources, while simultaneously ensuring that the information is tuned to the corpus. Caution must be taken with both these approaches, however, since similarity is often used as a basis for ontology creation. If this is the case, using the thesaurus as a means of calculating similarity may not be appropriate.

3. Ontology Creation and Tuning

It is unlikely that any pre-existing ontology will be entirely suitable for an NLP application, particularly within the area of terminology. Customisation of lexical resources to the domain is therefore an important task for NLP. Both term recognition and clustering can assist in this process, by structuring the knowledge acquired. Clustering is also useful for knowledge acquisition because information can be filtered from clusters down to individual lexical items. Experimental work has been carried out on clustering contexts using syntactic and semantic frames (Maynard and Ananiadou, 1999b). The contexts acquired previously in TRUCKS are segmented, and analysed, using the Supertagger tools developed by Joshi and Srinivas (1994). This allocates detailed syntactic information, including dependency relations, to chunks of texts. From these, a set of syntactic patterns is collected. This consists of 4 very general patterns:

- 1. NP + V
- $2. \quad V + NP$
- 3. NP + P + NP + V
- $4. \quad V + NP + P + NP$

After a normalisation process has been applied, these patterns are then mapped to semantic frames, using a set

of noun classes derived from the semantic tags from UMLS already acquired, and a set of verb classes acquired from WordNet. These are shown in Table 2. Noun tags are denoted in upper case, verb tags in lower case for ease of identification.

Class	Abbreviation	Examples
Body Part	<bp></bp>	corneal_epithelium
Abnormality	<ab></ab>	keratinous_debris
Phenomenon	<phen></phen>	spontaneous_rupture
Finding	<fin></fin>	ocular_haemorrhage
Virus	<vir></vir>	papilloma_virus
Injury	<inj></inj>	perforating_wound
Procedure	<proc></proc>	electron_microscopy
Body Substance	<bs></bs>	cholesterol_crystals
Section	<sec></sec>	lower_third
Size	<si></si>	circumference
Affect	<aff></aff>	change, affect
Description	<desc></desc>	occur, define
Observation	<obs></obs>	show, indicate
Position	<pos></pos>	occupy, envelop
Procedure	<proc></proc>	cut, remove

Table 2: Noun and verb classes

We give below an example of the whole procedure. A set of contexts is extracted from TRUCKS:

crosses cornea vitreous chamber traversed traverses tumour optic nerve surrounded envelops iris

After allocating syntactic and semantic categories, we get the following:

crosses cornea	V(act) + NP	<pos><bp></bp></pos>
vitreous_chamber traversed	NP + V(pas)	<bp><pos></pos></bp>
traverses tumour	V(act) + NP	<pos><bp></bp></pos>
optic_nerve surrounded	NP + V(pas)	<bp><pos></pos></bp>
envelops iris	V(act) + NP	<pos><bp></bp></pos>

After normalisation, and reordering, we produce the following semantic frames:

cross cornea
corneaconsectionchamber
consection

It is apparent from the final frames that the contexts have far more in common than they first appeared to. Since they have similar syntax and semantics, they can be clustered into one group. Table 3 depicts some examples of other clusters created. These clusters can be used,

³ personal communication, Dr. Malcolm Duncan

amongst other things, to resolve ambiguity. For example, if we come across a context of the form "X traverses vitreous_chamber", where X is an ambiguous term, we can predict that it will have the same meaning as X found in the context "X crosses the cornea" (Table 3). Although only a prototype system has been developed, it demonstrates the feasibility of a corpus-based approach for ontology creation and lexical tuning.

Pattern	Lexical String	
<obs><ab></ab></obs>	show inflammatory_changes	
	see atrophic_changes	
	show vascuolar_changes	
	see hyaline_changes	
	show degenerative_changes	
<bp><pos></pos></bp>	track cross	
	large_vessel enter	
	adventitia_overlie	
	vessel_enter	
<pos><sec>of<bp></bp></sec></pos>	traverse axis of cornea	
	envelop lower_half of iris	
	overlie surface of tumour	
	occupy portion of globe	

Table 3: Examples of clustered contexts

4. Using Thesauri for Evaluation Purposes

One of the major problems with the evaluation of terminological applications is that there is no gold standard against which their performance can be measured. If there existed a perfect set of terms for a domain, ATR systems would be unnecessary. Evaluation of term extraction methods is to a certain extent subjective, since it is dependent both on the application and on any criteria imposed by the end-user or evaluator.

For the evaluation of term extraction systems such as TRUCKS, two main alternatives exist: comparison with a manually created set of terms, and comparison with the terms found in some pre-defined list. We performed experiments with the TRUCKS system to compare two such methods of evaluation. Comparison of the set of terms listed in the UMLS and a set of terms extracted from the corpus and verified by domain experts showed an intersection of only 20%, both sets having omissions and false positives. The opinion of the experts is subjective and over-generative, but the UMLS is also incomplete.

Section	UMLS	Manual
Тор	13%	41%
Middle	8%	29%
Bottom	6%	27%

Table 4: Precision results with 2 evaluation methods

We also performed experiments comparing the results of the term extraction generated by TRUCKS against both UMLS and the list of terms verified by the experts. The top set of results (the uppermost third of the list of ranked candidate terms) was found to differ in precision by 28%, according to whether it was compared with UMLS or the manual list. Table 4 shows the difference in precision between the two methods for each section of the list of terms. Clearly the difference is a substantial one, showing an error rate of 68%.

5. Conclusions

For general language NLP applications, it may be that existing lexical resources are sufficient, but this is not the case for specialised domains. Current domain-specific thesauri and ontologies can provide a starting point for terminological work, but at the very least, they require lexical tuning. It is clear from our work that, while the UMLS may be the best option currently available, better lexical resources would improve results. Furthermore, problems arise from the use of resources such as the UMLS which have been designed primarily as information resources rather than linguistic ones. Using an information resource for a linguistic application is not always ideal, which leads to suggest the possibility of a merger of the two. Alternatively, we must turn to techniques involving the automatic derivation of resources from corpora, or at least, the tuning of existing resources to a domain or application. Such techniques are, however, still largely in their infancy, and evaluation needs to be carried out.

References

- Grefenstette, G. (1994). Explorations in Automatic Thesaurus Discovery, Kluwer Academic Publishers.
- Joshi, A. and Srinivas, B. (1994). Disambiguation of super parts of speech (or supertags): Almost parsing. In Proc. of 15th International Conference on Computational Linguistics (COLING), Kyoto, Japan.
- Lynch, C. (1997). Medical Terminology Management. In Handbook of Terminology Management, Wright, S.E. and Budin, G. (eds.), vol 1, pp. 160-170, John Benjamins, Amsterdam.
- Maynard, D.G. and Ananiadou, S. (1999a). Identifying Contextual Information for Term Extraction. In Proc. of 5th International Congress on Terminology and Knowledge Engineering (TKE '99), pp. 212-221, Innsbruck, Austria.
- Maynard, D.G. and Ananiadou, S. (1999b). A Linguistic Approach to Context Clustering. In Proc. of Natural Language Processing Pacific Rim Symposium (NLPRS), Beijing, China, pp. 346-351.
- Maynard, D.G. and Ananiadou, S. (1999c), Term Extraction using a Similarity-Based Approach. To

appear in "Recent Advances in Computational Terminology", John Benjamins.

- Mima, H., Ananiadou, S. and Tsujii, J-I. (1999). A Web-Based Integrated Knowledge Mining Aid System Using Term-Oriented Natural Language Processing. In Proc. of Natural Language Processing Pacific Rim Symposium (NLPRS), Beijing, China, pp. 13-18.
- National Library of Medicine. (1997). UMLS Knowledge Sources, 8th edition, U.S. Dept. of Health and Human Services, Bethesda, MD.
- Sager, J.-C. (1990). A Practical Course in Terminology Processing. John Benjamins.
- Smeaton, A. and Quigley, I. (1996). Experiments on using semantic distances between words in image caption retrieval. In Proc. of 19th International Conference on Research and Development in Information Retrieval, Zurich, Switzerland.
- Soderland, S., Fisher, D., Aseltine, J. and Lehnert,W. (1995). CRYSTAL: Inducing a Conceptual Dictionary. In Proc. of the 14th International Joint Conference on Artificial Intelligence (IJCAI-95).
- Ushioda, A. (1996). Hierarchical Clustering of Words. In Proc. of 16th International Conference on Computational Linguistics (COLING), pp. 1159-1162.
- Wilks, Y. and Stevenson, M. (1998). The Grammar of Sense: Using part-of-speech tags as a first step in semantic disambiguation. Journal of Natural Language Engineering 4(3).
- Yarowsky, D. (1992). Word sense disambiguation using statistical models of Roget's categories trained on large corpora. In Proc. of 14th International Conference on Computational Linguistics, pp. 454-460.