Term Sense Disambiguation Using a Domain-Specifc Thesaurus

Diana Maynard and Sophia Ananiadou

Dept. of Computing & Mathematics,
Manchester Metropolitan Universit y,
Manchester, M1 5GD,
UK

Abstract

Term extraction is important for many information systems applications. Although terms should be monoreferential, in reality they exhibit a high degree of ambiguity. Whilst conventional solutions mainly involve statistical approaches, this paper proposes a more hybrid technique. The approach is based on the identification of relevant contextual information from sublanguage texts, enabling the incorporation of both syntactic and semantic knowledge in addition to statistical information. Our approach uses a linguistic filter to identify word classes which are expected to convey salient information about the term, and a semantics-based matching technique to find terms occurring in similar environments. Emphasis is placed on the **relationship** between a term and its context words, rather than simply relying on statistical information about frequency of occurrence. Semantic information is acquired from UMLS, a domain-specific electronic medical thesaurus, and the context words are classified according to their significance in relation to the term.

1. Introduction

The increasing volumes of electronic textual data and the development of new applications have given rise to a variety of disambiguation techniques. Term ambiguity can greatly reduce the efficiency of term extraction applications such as information extraction and retrieval, dictionary construction, automatic indexing, machine translation and hypertext linking.

Traditional methods of sense disambiguation involve comparing the distribution of the words in a context window with typical distributions of the word's senses to find the closest match (Gale et al., 1992). The use of lexical semantic information has largely been ignored for term disambiguation, although it is increasingly being used for other NLP applications. Term sense disambiguation presupposes a method of term extraction, and we thus build on a tool for automatic term recognition developed in (Frantzi and Ananiadou, 1997), which uses a combination of linguistic and statistical information acquired from the corpus.

2. Ambiguity in Terms

Terms are theoretically the embodiment of specialised concepts (Sager, 1990). Ideally, there should therefore be a univocal relationship between a term and a concept, but in practice, terms exhibit a high degree of both polysemy and synonymy. The level of ambiguity can be reduced substantially by restricting the domain to a sublanguage, but ambiguous terms are by no means eliminated. For example, in the medical domain the term *drug* found in a medical text could refer to either an illegal substance or simply to a kind of medicine. Similarly, there is nothing to prevent the general language meaning of a term occurring in a specialised text. This is partly because terms are often created by analogy to general language words. This is particularly the case in newer fields such as computing, with terms like *mouse*, *window* and *desktop*. According to Cimino et al.(Cimino et al., 1989), the meaning of a concept should be defined by its position and relationship to other concepts in a network. Again, in practice this is not generally the case, as almost all hierarchies still suffer from the problem of multidimensionality, whereby the same concept can be classified in more than one way.

2.1. Types of Ambiguity

Some ambiguous terms are simply homographs, e.g. *tear*. This clearly has two unrelated meanings and represents two different concepts (a "rip" and "ocular fluid"). Other terms can be considered as hyponyms of different terms, and thus be represented in different parts of the hierarchy, even though they only refer to a single concept. For example, the term *hydroxycalciferol* only represents one concept, but can be a hyponym of either *steroid* or *vitamin*, and thus be assigned two different positions in the network.

Although multidimensionality is an inherent problem of hierarchical networks, it is really the result of a more cognitive issue - the way in which we view the world. For example, the term *football boot* can be seen as either an item of footwear or as a piece of sports equipment. Under which category should it therefore be placed in a hierarchical representation? Most probably, it would be placed under both, or there would be a link from one position to the other. How the term is best categorised might depend on the individual circumstance. For example, in a text about football it might be more useful to know that a football boot is an item of footwear; in a text about footwear it might be more useful to know that it is a piece of sports equipment, since this provides more specific information about the term. Since it is virtually impossible to avoid a multidimensional representation, we need to find some means of accessing the term in its most relevant location according to the situation.

In many ways, term sense disambiguation is similar to the problem of word sense disambiguation. However, it dif-

fers in two main aspects. Firstly, terms are domain-specific, which means that general language resources and techniques may not be appropriate. Secondly, the majority of technical terms consist of more than one word. The ambiguity of multiword terms is generally not caused by different senses of the individual components of the term, but by different senses of the term as a whole. The different meanings of the term may be linked to different domains, but they may equally be present within a specific domain. Although domain-specific text is likely to contain the meanings of a term related to that domain, it does not rule out the possibility of the general meaning of that term being used as well.

3. Related Work

Previous approaches to the extraction of semantic information have tended to use general language dictionaries or thesauri. One common method is to compare the content of the dictionary definition of a word with the words in the surrounding context. (Lesk, 1986) used an online dictionary to calculate the overlap between dictionary definitions of word senses. (Smeaton and Quigley, 1996) calculated word-word similarity between related words in a thesaurus. (Yarowsky, 1992) captured the meanings of word senses from semantic categories in Roget's Thesaurus.

The application of these methods to word sense disambiguation has been fairly successful. There are, however, two main reasons why they are not entirely suitable for our work. Firstly, they are used for word sense rather than term sense disambiguation. Secondly, they are designed to deal with general rather than domain-specific language. Technical terms are not covered sufficiently in any general language dictionary or thesaurus, and we propose that they really require a more specialised source of information.

An alternative approach to the acquisition of semantic information involves the use of a disambiguated corpus for training. (Riloff and Lehnert, 1992) developed an algorithm to derive relevancy cues from training texts, which they used for information extraction. (Soderland et al., 1995) also developed a system to identify concepts in a text, by means of linguistic features which reliably identified the conceptual content of a phrase. (Grefenstette, 1994) adopted a hybrid, knowledge-poor approach to the automatic extraction of semantic information from large corpora, using only syntactic information. The advantage of these techniques is that they require no dictionary or thesaurus and provide information specific to the corpus. However, they do require suitable large-scale corpora.

Linguistic information has been used quite extensively in multi-word term extraction, e.g. (Daille et al., 1994), (Bourigault, 1992) (Justeson and Katz, 1996), but it consists mainly of shallow syntactic information in the form of a grammar restricting the types of phrases to be extracted as potential terms. (Lauriston, 1996) found that introducing semantic features from WordNet did not significantly improve term recognition, but points out that this may have been due to technical rather than fundamental reasons.

4. The Use of Contextual Information

4.1. NC-Value method

A method for term sense disambiguation relies on having a suitable means of extracting terms. Our approach uses a method for multi-word automatic term recognition called NC-value, developed by (Frantzi and Ananiadou, 1997). This uses contextual information to improve the ranking of candidate terms, and is thus suitable for our purposes since we can make use of the knowledge already acquired, and extend it by incorporating deeper forms of linguistic information. Terms do not appear arbitrarily in sentences, but are related syntactically and semantically to their context, so we should be able to make use of this information both for recognising and disambiguation terms.

In Frantzi's approach, potential terms are first extracted from a corpus and ranked using the C-value method (Frantzi and Ananiadou, 1996) based on frequency of occurrence and term length. Contextual information is then incorporated into the algorithm, in the form of weights based on the statistical characteristics of the context words.

Context words are composed of any nouns, adjectives and verbs in a fixed-size context window, either preceding or following a potential term. These context words are extracted and assigned a weight based on how frequently they appear with terms.

$$Weight(w) = 0.5(\frac{t(w)}{n} + \frac{ft(w)}{f(w)})$$

where
w is the context word
n is the number of top-ranked candidate terms considered
t(w) is the number of candidate terms w appears with
ft(w) is the total frequency with which w appears with candidate terms
f(w is the total frequency of w in the corpus.

The NC-value combines the original C-value with this weight to produce a reranking of the list of terms.

4.2. Restrictions on the Context

The type of context used for the extraction of terms is also an important issue. Although a domain-specific corpus is used, it is clear that not all the words in the context are equally useful. Frantzi's use of contextual information stems from the idea that multi-word terms are less easily modifiable than non-terms (Sager, 1978), so information about the modifiers can be used to evaluate potential terms. Frantzi found that not only adjectives and nouns, but also verbs are potential sources of information. Although it is possible that other parts of speech could play a role, it is less evident how the meaning of closed syntactic categories (such as adverbs or prepositions) could be incorporated.

Frequency-based methods do not generally take into account the different syntactic categories of context words, nor their position within the context window. One exception to this is Radford et al.'s dynamic matching algorithm (Radford et al., 1997), which applies structural constraints

to the word order of the context window when calculating similarity between contexts. It does not, however, consider the position of the term in relation to the context word.

For our purposes, this position is considered important. For example, in English, adjectives preceding a term are more likely to be relevant than adjectives following a term, with the exception of predicative adjectives. The position of the verb could also be of importance, since it indicates different semantic roles for the term. The significance of the position of another noun is less apparent.

4.3. Incorporating Linguistic Information

The linguistic knowledge used in the NC-value approach is very limited, since it only takes into account the syntactic category of the context word. It considers neither any differences that might exist between the categories, nor the position in which the context word occurs with respect to the term. For example, it might be the case that nouns are more useful than verbs in predicting termhood, or that verbs preceding a term are more useful indicators than verbs following a term.

Another factor which plays an important role is that of the *meaning* of the context words. We predict that words which are closely related to a potential term are more likely to indicate termhood than those which are unrelated. Also, context words which are themselves terms should be an indicator. We propose to incorporate these factors into the weighting system to improve the ranking of the terms.

Our claim is that if a context word has some contribution towards the determination of a term, there should be some significant correspondence between the meaning of that context word and the meaning of the term. In other words, there should be some identifiable semantic relation between the two, which can be used to make a contribution towards the correct identification and comprehension of a term.

5. Uncovering Semantic Relations

The context word and term are first generalised into their semantic types, using information provided by the UMLS Metathesaurus. Generalisations can then be made about these semantic types and the relations between them. The context words are divided into those that themselves are potential terms (i.e. those that are contained in the UMLS Metathesaurus) and those that are not. Just because a word is part of the context surrounding a term does not prevent it from being a term itself – in fact it is highly likely considering the high proportion of nouns in a technical text that are terms. Context words which are terms will automatically receive a higher weighting, since they are likely to have a stronger association with other terms than non-term context words are.

Initially, it is these potential terms which are examined as context words. They are first grouped together according to their position in the semantic network, and then their relationship with the candidate terms is determined. This is achieved in two ways: by measuring the similarity between the two, and by examining the relationships between them in the UMLS Semantic Network. Using this information, we aim to assess the contribution of different types of context words to the understanding of the term.

5.1. Semantic Information in UMLS

The UMLS (Unified Medical Language System) (NLM, 1997) is a set of knowledge sources containing information about medical terminology, organised in a hierarchical structure. Not only does it provide a classification system for the terms, but it also contains morphological, syntactic and semantic information. The Semantic Network contains additional information about the relations between the semantic classes to which the terms are assigned.

Each term and context word is tagged with its semantic type from the UMLS Metathesaurus. For example, the term *actinic keratosis* will be tagged with the semantic type *acquired abnormality*.

Relational information about the semantic types is also available. This information is generic rather than specific, i.e. details are provided about classes of terms rather than about individual terms, and do not necessarily hold for every member of that class. For example, a relationship exists between the semantic classes of *disease or syndrome* and *acquired abnormality*:

```
DISEASE OR SYNDROME|RESULT OF|
ACQUIRED ABNORMALITY
```

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.

5.2. Calculating Similarity

The similarity between context word and term is measured using the UML Semantic Network. Our approach is modelled on techniques used in Example-Based Machine Translation (EBMT) (Zhao, 1996), (Sumita and Iida, 1991). EBMT aims to select the most similar example for a given problem, by defining a set of relevant examples and comparing them with the problem in terms of context and configuration. Most of these comparison methods involve using some kind of semantic distance measure, based on the relative positions of the two items within a hierarchical network.

In its simplest form, similarity is measured by edge-counting – the shorter the distance between the words, the greater their similarity. The Most Specific Common Abstraction (MSCA) method (Kolodner, 1993), which involves tracing the respective paths of the two words back up the hierarchy until a common ancestor is found, and then measuring the average distance from node to MSCA. The shorter the distance to the MSCA, the more similar the two words. The position of the MSCA within the hierarchy can also be used to measure similarity (Zhao, 1996), (Sumita and Iida, 1991). The lower down in the hierarchy the MSCA, the more specific it is and therefore the more information is shared by the two concepts, thus making them more similar. Both Sumita & Iida and Zhao combine this "positional

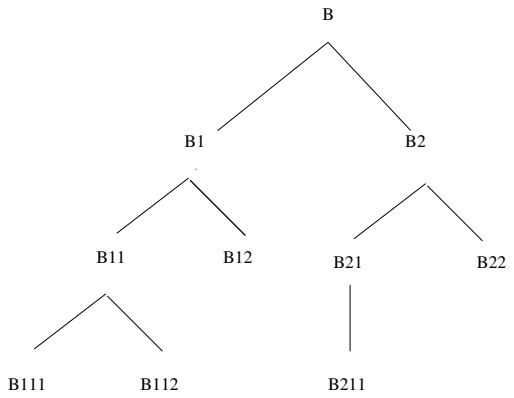

Figure 1: Fragment of the Semantic Network

weight" with an "importance weight" based on the relative importance of the attribute being considered. We combine the positional weight with a "commonality weight" which takes into account the number of common nodes shared by the two words. The nodes in the Semantic Network are coded such that the number of digits in the code represents the number of leaves descended from the root to that node, as shown in Figure 1. Similarity between two nodes is calculated in a very simple but effective way: descending from the root node to each of the nodes in question, the commonality weight is the number of nodes in common traversed, whilst the positional weight is the total number of nodes traversed. The commonality weight is then divided by the positional weight to produce a similarity figure between 0 and 1 (1 being the case where the two nodes are identical).

$$sim(w1, w2) = \frac{com(w1, w2)}{pos(w1, w2)}$$

where
com(w1,w2) is the number of nodes shared by word 1 and word2
pos(w1,w2) is the total number of nodes traversed between the root node and word 1 and word2

So, taking the hierarchy depicted in Figure 1, the similarity between nodes B1 and B2 would be calculated as follows:
com(w1,w2) = no. of nodes in common = 2
pos(w1,w2) = total no. of nodes = 4
sim(w1,w2) = 2/4 = 0.5
In the same way, the similarity between nodes B111 and B112 would be:
com(w1,w2) = 6
pos(w1,w2) = 8
sim(w1,w2) = 6/8 = 0.75
It becomes clear that the further down in the hierarchy the two items are, the greater their similarity, which is intuitive since they are more specific: there is greater similarity betwen *apple* and *banana* than between *fruit* and *vegetable*.

It is also intuitive that the greater the horizontal and/or vertical distance between words in the network, the less similar they are. For example, the similarity between nodes B and B112 would be 0.4, whilst the similarity between nodes B111 and B211 would be 0.25.

5.3. Relations between Semantic Types

Although the degree of similarity between context word and term acts as a useful indicator of the significance of that context word, it is not completely reliable. On investigating the UMLS Semantic Network, it becomes clear that there are many significant relationships between words which are not similar in terms of their semantic types. For example, the semantic types *acquired abnormality* and *mammal* only have a similarity of 0.44, but there is a relationship between the two in that acquired abnormalities *affect* mammals. This relationship can be seen as an indicator that if, for example, we find a context word belonging to the semantic type *acquired abnormality*, it is relevant to a term in its vicinity belonging to the semantic type *mammal*, even if the two words themselves may not appear to be strongly related semantically.

6. Conclusions and Further Work

In this paper we have presented a semantic approach to term sense disambiguation. Information about the context is acquired from a domain-specific thesaurus, and relationships between the context and term are defined in terms of both similarity of meaning and hierarchical relations. A simple similarity measure based on a hierarchical network enables context words to be grouped together according to their meaning and compared with the terms with which they appear. Relational information from the thesaurus also enables the relevance of the context word to the term to be assessed.

Work still remains to be done on this latter part, so that weights can be correctly assigned to context words. Once this has been carried out, enhancements can be made to the term extraction process, in the form of a reranking of the list of candidate terms, and to the disambiguation procedure. More general linguistic information about basic semantic roles could also be added to augment the weights.

Currently, the context words which are not terms have not been considered. It is clear that these belong to a fairly restricted subset, and thus their relationship with the terms should not be ignored. In particular, we would like to examine the role of verbs as context words, since it has already been shown (Frantzi and Ananiadou, 1997) that these play a significant role in the prediction of terms.

The approach adopted is domain-specific and knowledge-rich, although the techniques used are portable to other domains. Using a sublanguage corpus ensures a higher concentration of terms, and imposes greater restrictions at both semantic and syntactic levels. Whilst this narrows the interpretation of linguistic patterns and reduces ambiguity to a certain degree, domain-specific resources are still necessary since general language resources do not contain enough specialised information. Although knowlege-poor approaches have been successful for general language corpora, we propose that the incorporation of semantic knowledge is necessary for term sense disambiguation in sublanguages.

7. References

8. *

References

Bourigault, D. (1992). Surface grammatical analysis for the extraction of terminological noun phrases. In *Proc. of the 14th International Conference on Computational Linguistics*, pages 977–981.

Cimino, J., Hripcsak, G., Johnson, S., and Clayton, P. (1989). Designing an introspective, controlled medical vocabulary. In Kingsland, L., editor, *Proc. of 14th Annual Symposium on Computer Applications in Medical Care*, pages 513–518. IEEE Computer Society Press.

Daille, B., Gaussier, E., and Langé, J. (1994). Towards automatic extraction of monolingual and bilingual terminology. In *Proc. of COLING 94*, pages 515–521.

Frantzi, K. and Ananiadou, S. (1996). A hybrid approach to term recognition. In *Proceedings of NLP+IA 96*, volume 1, pages 93–98, Moncton, Canada.

Frantzi, K. and Ananiadou, S. (1997). Automatic term recognition using contextual cues. In *Proceedings of 3rd DELOS Workshop*.

Gale, W., Church, K., and Yarowsky, D. (1992). Work on statistical methods for word sense disambiguation. In *Probabilistic Approaches to Natural Language: Papers from the 1992 Fall Symposium*, pages 54–60, Cambridge, Massachussetts. AAAI Press.

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

Justeson, J. and Katz, S. (1996). Technical terminology: some linguistic properties and an algorithm for identification in text. *Natural Language Engineering*, 3(2):259–289.

Kolodner, J. (1993). *Case-Based Reasoning*. Morgan Kaufmann Publishers Inc., San Mateo, California.

Lauriston, A. (1996). *Automatic Term Recognition: performance of Linguistic and Statistical Techniques*. PhD thesis, UMIST, Manchester.

Lesk, M. (1986). Automatic sense disambiguation: how to tell a pine cone from an ice cream cone. In *Proc. of the SIGDOC Conference*, pages 24–26.

NLM (1997). *UMLS Knowledge Sources*. U.S. Dept. of Health and Human Services, 8th edition.

Radford, I., Ananiadou, S., Arranz, M. V., and Tsujii, J. (1997). Term disambiguation by adding structural constraints to lexically based context matching techniques. In *NLPRS97*.

Riloff, E. and Lehnert, W. (1992). Classifying texts using relevancy signatures. In *Proc. of AAAI*.

Sager, J. (1990). *A Practical Course in Terminology Processing*. John Benjamins.

Sager, J. C. (1978). Commentary. In *Table Ronde sur les Problémes du Décourage du Terme*, pages 39–52. Service des Publications, Direction des Française, Montréal.

Smeaton, A. and Quigley, I. (1996). Experiments on using semantic distances between words in image caption retrieval. In *Proc. of the 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*.

Sumita, E. and Iida, H. (1991). Experiments and prospects of example-based machine translation. In *Proc. of 29th Annual Meeting of the Association for Computational Linguistics*, pages 185–192, Berkeley, California.

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*, pages 454–460.

Zhao, G. (1996). *Analogical Translator: Experience-Guided Transfer in Machine Translation*. PhD thesis, Dept. of Language Engineering, UMIST, Manchester, England.