Acquiring Contextual Information for Term Disambiguation

Diana Maynard and Sophia Ananiadou
Dept. of Computing & Mathematics
Manchester Metropolitan University
Manchester, M1 5GD, U.K.
{D.Maynard, S.Ananiadou}@doc.mmu.ac.uk

Abstract

Term extraction is important for many information systems applications. Although terms should be monoreferential, they actually exhibit a high degree of ambiguity. This paper describes a method for acquiring contextual information for term disambiguation. Unlike traditional statistical techniques, semantic information about the context is acquired from a domain-specific thesaurus, in order to define relationships between context and term. Using this information, we evaluate the significance of context words with respect to their co-occurring terms.

1 Introduction

The increasing volumes of electronic textual data in recent years have given rise to a need for term extraction applications, for example in dictionary construction, automatic indexing, machine translation, information retrieval and hypertext linking. The fact that terms exhibit a high degree of ambiguity can be problematic for these applications.

Traditional methods of word sense disambiguation rely on statistical methods, and 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). Although the use of lexical semantic information is increasingly being used for other NLP applications, it has largely been ignored for term disambiguation.

In this paper, we examine 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 semantic similarity with the co-occurring term. Context words are separated into those which are themselves terms and those which are not, and an experiment is performed to determine the different contributions each play towards term extraction. By gaining a better understanding of the relationship between term and context, we aim to improve the processes of both term extraction and disambiguation.

2 Term sense disambiguation

In many ways, term sense disambiguation is similar to the problem of word sense disambiguation. However, it differs 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. For example, in biochemistry, the term *complement* is used to refer to a component of blood serum, but it might equally be found in a medical text with its general meaning. Similarly, the term *drug* found in a medical text could refer to either an illegal substance or simply to a kind of medicine.

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. (Bourigault, 1992), (Daille et al., 1994), (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.

4 Acquiring Contextual Information

4.1 NC-Value method

A method for term sense disambiguation relies on having a suitable means of extracting terms initially. 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. 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, and these weights are combined to produce a *context factor* for each term. The NC-value combines the original C-value with this context factor to produce a reranking of the list of terms:

$$NCvalue(a) = 0.8 * Cvalue(a) + 0.2 * CF(a) \quad (1)$$

where
a is the candidate term,
$Cvalue(a)$ is the Cvalue for the candidate term,
$CF(a)$ is the context factor for the candidate term.

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.

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 propose two new indicators of termhood:

- context words which are themselves terms (which we call *context terms*)
- context words which are closely related in meaning to their co-occurring 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. In the following two sections we outline how this is achieved.

5 Uncovering Semantic Relations

Using a million-word corpus of eye pathology records, candidate terms are extracted using the NC value, along with context windows of 5 words either side of the term. Each context word and term is then tagged with its semantic type, 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 predicted to have a stronger association with other terms than non-term context words have.

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 by measuring the similarity between the two in the UMLS Semantic Network. Using this information, we aim to assess the contribution of different types of context words with respect to 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 term 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. We plan to incorporate this relational information into our future work.

5.2 Calculating Similarity in the UMLS Semantic Network

The similarity between context term and term is measured using the UMLS 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) 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. We define two weights:

- *positional*: the vertical position of the nodes within the hierarchy, measured by the combined number of nodes belonging to each word
- *commonality*: essentially the horizontal distance between the two nodes, measured by the number of shared common ancestors multiplied by the number of words (usually two).

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

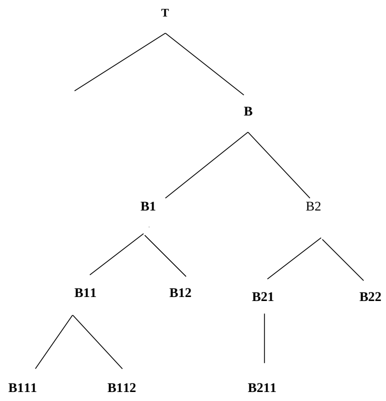

Figure 1: Fragment of the Semantic Network

to that node, as shown in Figure 1 (where the root node is labelled T). Similarity between two nodes is calculated by dividing the commonality weight by the positional weight to produce a figure between 0 and 1, 1 being the case where the two nodes are identical, and 0 being the case where there is no common ancestor.

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

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 word2.

So, taking the hierarchy depicted in Figure 1, the similarity between nodes B111 and B112 would be calculated as follows:-

com(B111, B112) = 3 * 2 = 6
pos(w1,w2) = 4 + 4 = 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: for example there would be greater similarity between *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.

6 Incorporating Similarity into the NC Value

6.1 Calculating the Similarity Weight

The similarity measure described in the previous section measures the similarity between each term and the "significant" context words it occurs with. The similarity weight is then added into the original NC value in the following way.

The original NC value first calculated a weight for each context word appearing with a top-ranked term, and then a context factor for each term (see section 4.1). The similarity weights are incorporated into the context factor in the same way as the weights for the context words, i.e. for each context term appearing with a candidate term, the similarity weight is multiplied by the frequency of the context term appearing with the candidate term. This is added to the original context factor. More formally, the whole context factor can be described as:

$$CF(a) = \sum_{b \epsilon C_a} f_a(b) \cdot weight(b) + \sum_{d \epsilon T_a} f_a(d) \cdot sim(d, a) \quad (2)$$

where
a is the candidate term,
$C_a$is the set of context words of a,
b is a word from C_a,
$f_a(b)$ is the frequency of b as a context word of a,
$weight(b)$ is the weight of b as a context word,
T_a is the set of context terms of a,
d is a word from T_a,
$f_a(d)$ is the frequency of d as a context term of a,
$sim(d, a)$ is the similarity weight of d occurring with a.

The procedure for incorporating the similarity weights into the NC value is therefore as follows:

- For each term, find a context window of 5 words either side which are adjectives, nouns or verbs.
- From this context window, extract all the terms to produce pairings of term and context term.
- Find the hierarchical position of each from UMLS.
- Calculate the similarity between each pair (see section 5.2).
- Calculate the new context factor for the term.
- Recalculate the NC value replacing the old context factor with the new one (see section 4.1).

7 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 semantic relationships between the context and term are defined. 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.

By incorporating semantic information into the NC Value measure, we attempt to distinguish context words according to their importance as a means of predicting termhood. Currently work is being carried out to evaluate the different contributions made by context terms and other context words, but it is already clear that the semantic information obtained from context terms has a valuable role in its association with co-occurring terms.

Currently, the semantics of context words which are not terms has 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, 1997) that these are closely linked to terms.

Our approach attempts to provide some useful developments to both term extraction and disambiguation techniques. The method is domain-specific and knowledge-rich, although the techniques used are portable to other domains. Using a sublanguage corpus imposes greater syntactic and semantic restrictions on the text, but domain-specific resources are still necessary since general language dictionaries and thesauri do not contain enough specialised knowledge.

References

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

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

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

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

K.T. Frantzi. 1997. Incorporating context information for the extraction of terms. In *Proceedings of ACL/EACL '97*, pages 501–503, Madrid, Spain, July.

W.A. Gale, K.W. Church, and D. Yarowsky. 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.

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

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

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

M. Lesk. 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, U.S. Dept. of Health and Human Services, 1997. *UMLS Knowledge Sources*, 8th edition, January.

Ian Radford, Sophia Ananiadou, M. Victoria Arranz, and Jun'ichi Tsujii. 1997. Term disambiguation by adding structural constraints to lexically based context matching techniques. In *NLPRS97*.

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

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

A. Smeaton and I. Quigley. 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.

S. Soderland, D. Fisher, J. Aseltine, and W. Lehnert. 1995. CRYSTAL: Inducing a conceptual dictionary. In *Proc. of the 14th International Joint Conference on Artificial Intelligence*.

E. Sumita and H. Iida. 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.

D. Yarowsky. 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.

Gang Zhao. 1996. *Analogical Translator: Experience-Guided Transfer in Machine Translation*. Ph.D. thesis, Dept. of Language Engineering, UMIST, Manchester, England.