TRUCKS: a model for automatic multi-word term recognition

Diana Maynard, Sheffield University, U.K.
Sophia Ananiadou, University of Salford, U.K.

March 30, 2000

Abstract

This paper examines the use of linguistic techniques in the area of automatic term recognition. It describes the TRUCKS model, which makes use of different types of contextual information - syntactic, semantic, terminological and statistical - seeking particularly to identify those parts of the context which are most relevant to terms. From an initial corpus of sublanguage texts, this identifies, disambiguates and ranks candidate terms. The system is evaluated with respect to the statistical approach on which it is built, and with respect to its expected theoretical performance. We show that by using deeper forms of contextual information, we can improve on the extraction of multi-word terms. The resulting list of ranked terms is shown to improve on that produced by traditional methods, in terms of precision and distribution, while the information acquired in the process can also be used for a variety of other applications, such as disambiguation, lexical tuning and term clustering.

1 Introduction

Increasing volumes of electronic text have given rise to a need for new methods of collecting, organising, storing and retrieving terminological data. Coupled with the laborious nature of such tasks, and the constant dynamism of technical fields, applications such as term extraction and disambiguation benefit greatly from automisation.

Methods for automising terminological procedures have traditionally been statistical, but these are not always sufficiently accurate, particularly for smaller corpora. Recently, hybrid approaches incorporating linguistic information have been developed, but these still rely largely on statistical information and use only very shallow syntactic knowledge. Approaches to word sense disambiguation and machine translation have taken advantage of contextual information in a more meaningful way, but terminology has rarely followed suit. We present an approach to term recognition which identifies salient parts of the context and measures their strength of association to relevant candidate terms.

Although statistical approaches to automatic term recognition, e.g. [3, 7, 11, 18, 19], have achieved relative success over the years, the addition of suitable linguistic information has the potential to enhance results still further, particularly in the case of small corpora or very specialised domains, where statistical information may not be so accurate. One of the main reasons for the current lack of diversity in approaches to term recognition lies in the difficulty of extracting suitable semantic information from specialised corpora, particularly in view of the lack of appropriate linguistic resources. The increasing development of electronic lexical resources, coupled with new methods for automatically creating and fine-tuning them from corpora, has begun to pave the way for a more dominant appearance of natural language processing techniques in the field of terminology.

The TRUCKS approach to term recognition (Term Recognition Using Combined Knowledge Sources) focuses on identifying relevant contextual information from a variety of sources, in order to enhance traditional statistical techniques of term recognition. We introduce an idea called **terminological acquaintance**, which considers terms in conjunction with the contextual information surrounding them, rather than in isolation. This exerts an influence on the way in which they are treated. Instead of being manipulated as statistical objects, a deeper understanding of their textual situation and relevance is acquired. Although contextual information has been previously used, e.g. in general language [16] and in the NC-Value method for term recognition [13, 15], only shallow syntactic information is used in these cases.

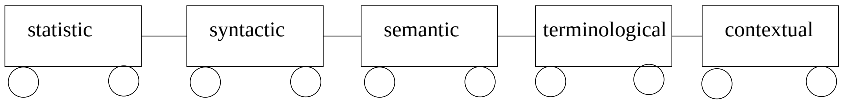

Figure 1: The TRUCKS Model

The model is called TRUCKS not only because it is a suitable acronym, but also because it evokes the structure of the system. This can be viewed as an engine pulling a set of trucks, each containing a particular type of knowledge (such as semantic, terminological, statistical and contextual). It is knowledge-rich in that it combines information from different sources: an external thesaurus and semantic network providing semantic and classificatory information, and the corpus providing statistical and syntactic information. The "trucks", or modules, are semi-independent in that each can be altered without affecting the others, but each plays a role in the assignment of weights which are used to determine the status of the candidate terms and their ranking.

2 Term Recognition and Disambiguation

2.1 Terms and Concepts

From a philosophical point of view, a term can be distinguished from a word (or from another linguistic "sign") in that its "semantic extension is defined in relation to the concept rather than the name" [31]. In other words, whilst lexicography is interested in the meaning of a lexical unit, terminology is interested in the concept represented by the lexical unit. Rondeau also points out that a concept is always defined in relation to other concepts within a particular field, so that terms should not be considered in isolation but as part of a larger group. Whilst words can also be classified in hierarchies or ontologies, this is not crucial to their meaning in the same way as it is with terms.

The correct association of term and concept is important for almost any application of computational linguistics involving specialised languages, just as it is for machine translation [10]. We therefore consider that an understanding of the relationship between term and concept is necessary in order to understand term variation and ambiguity, both of which are crucial to the theoretical study of terminology and its practical applications [20].

Ideally, terms should be monoreferential, i.e. there should be a one-to-one correspondence between term and concept [17]. In practice, however, this is not the case, and terms may exhibit both polysemy and synonymy. It is now generally acknowledged that monoreferentiality of terms is no longer practical or even necessarily desirable [32, 4]. The problem of term ambiguity is therefore strongly correlated with the problem of defining the relationship between term and concept. A term is ambiguous if it can represent more than one concept: for example, *plant* can refer to a vegetable, the sole of the foot, a young oyster, a swindle, or a set of machinery [25]. Conversely, terms are considered synonymous if they relate to a single concept: for example, *peewit*, *dastard plover*, *lapwing* and *wype* all refer to the same species of bird in zoology.

2.2 Term Ambiguity

The level of ambiguity can be reduced by restricting the domain to a sublanguage, but ambiguity can still exist within a single domain. For example, the term *drug* can refer equally to a prescribed medicine or to an illegal narcotic, both of which might appear in the medical domain. In the field of medicine in particular, multidimensionality is in strong evidence [9]. This occurs when a term is classified in more than one place in a hierarchy, and is more or less inevitable. For example, *pulmonary tuberculosis* is both a respiratory disease and an infectious disorder, and might therefore be represented in two different sections of the hierarchy. It can also occur where there is a kind of ambiguity in the term itself. For example, *coffee* can refer to the whole bean, or to the drink made from the bean; *television* can refer to the physical hardware or to the broadcast. In this case, both instances of the term belong to the same section of the hierarchy, but one might be represented at a higher node than the other.

Ambiguity is a particular problem in the field of medicine, because there can be very serious, even fatal, consequences if communication is not clear. The field is also especially susceptible to ambiguity because it is both multi-faceted and dynamic [32]. Blois claims that the problem is partly due to the "vertical" structure of the field, in that it ranges from very precise subdomains such as physics and chemistry to more general ones such as sociology and psychology [2]. Not only does this mean that the likelihood of ambiguity increases as soon as the more precise areas at the bottom of the scale are left behind, but it also has more far-reaching implications. Because knowledge from a wide range of fields is necessary, problems associated with multidimensionality and with inter-domain meaning change are amplified. A further problem is caused not by imprecision, but by terminological changes evoked by time. Just as general language evolves over time, due to a variety of social, cultural and historical factors, it would be surprising if terminology did not also change over time, given the dynamic nature of scientific fields such as medicine.

While much work has been carried out in the area of word sense disambiguation, the subject of ambiguity in terms has been largely avoided. There are two main reasons for this: firstly, because it is considered unimportant or not sufficiently widespread, and secondly, because it is deemed too difficult to tackle. The first may be true for certain applications: for example, most term extraction systems are not concerned with a term's meaning, but only its occurrence. Similarly, some applications may consider the restrictions of the domain sufficient to perform coarse-grained disambiguation, which is all they require. Finer-grained disambiguation may, however, be necessary for certain tasks such as creation and tuning of ontologies. The second reason, that term disambiguation is too difficult, stems from the fact that most approaches to terminology do not involve understanding, and therefore would necessitate much adaptation to cater for dealing with ambiguity.

2.3 Term Extraction

Approaches to term extraction can be roughly divided into 3 categories: linguistic, e.g. [1, 3], statistical, e.g. [33, 19] and hybrid, e.g. [6, 18, 15]. Closely related to term extraction methods are approaches to the recognition of lexical atoms and collocations, which tend to use similar methods [12, 27]. Multi-word term extraction approaches consist almost entirely of hybrid approaches similar to these, since single-track approaches are largely insufficient in this case, if they are not to be restricted to very narrow applications. Independently from this, the knowledge used to identify terms may be intrinsic or extrinsic to the term. Intrinsic knowledge might involve morphological, syntactic, semantic or statistical information about the individual term or its components (lexical items or even morphemes). Extrinsic knowledge makes use of contextual information (either linguistic or statistical) or information from exterior resources such as dictionaries, thesauri, ontologies, or additional corpora.

2.4 The TRUCKS Model

The method we propose for term extraction is drawn from a combination of approaches, using information from a variety of sources, both intrinsic and extrinsic, and combining linguistic and statistical knowledge. By compiling a set of weights for each term, dependent on the relative importance of these different types of knowledge, candidate terms are selected and ranked in order of termhood. The model consists of three layers:

1. a base statistical layer, determined by the C-Value [14], selects candidate terms from the text.
2. a middle layer, determined by a Context Weight [15], combines linguistic and statistical information about context words.
3. an upper layer, determined by an Importance Weight [24], considers syntactic, semantic and terminological information about context words.

These three weights are combined to form the SNC-Value, which is described more fully in the following 2 sections.

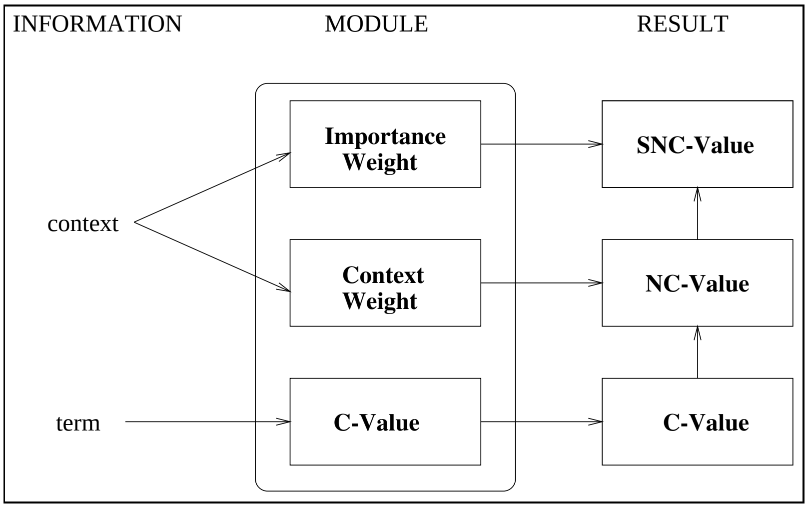

Figure 2: Architecture of the TRUCKS Model

The architecture of the system is depicted in Figure 2. From this, we can see that the system can be used at any of its three levels, but that each level is dependent on the levels beneath it. For example, the NC-Value is dependent on the C-Value and the context weight. If changes are made to either of these, the NC-Value result will be affected. However, changes to the C-Value do not affect the context weight. Similarly, changes to the C-Value affect the SNC-Value, but changes to the Importance Weight do not affect the NC- or C-Values. The C-Value operates solely on terms, whereas the Context Weight and Importance Weight operate on contextual information as well.

3 Contextual Information: a Term's Family and Friends

Just as a person's social life can provide valuable clues about their personality, so we can gather much information about the nature of a term by investigating the company it keeps. Little use of contextual information has previously been made for term extraction other than from a syntactic point of view. This generally takes one of two forms: barrier word methods and co-occurrence methods. Barrier word methods, e.g. [3, 28] are based on the principle of identifying boundaries demarcating candidate terms, usually according to their syntactic properties. Co-occurrence methods such as the NC-Value method [13], are based on the principle that terms are restricted in the ways that they can be modified, and that particular words in the context therefore occur more frequently with terms than with non-terms. Deeper forms of contextual information have largely been ignored for term extraction, but have been used for tasks such as the identification of synonyms [16] and the identification of terminological relations [8, 26].

In TRUCKS, we extract three different types of contextual information to form the top layer of the model, the Importance Weight:

1. syntactic;
2. terminological;
3. semantic.

	Non-Terms		Terms	
Category	Total	%	Total	%
Verb	13147	28	16170	32
Prep	12050	29	15551	31
Noun	11500	26	12425	24
Adj	7682	17	6604	13
Total	44379	100	50750	100

Table 1: Distribution of boundary words

Category	Weight
Verb	1.2
Prep	1.1
Noun	0.9
Adj	0.7

Table 2: Weights for categories of boundary words

3.1 Syntactic knowledge

Syntactic knowledge is based on words in the context which occur immediately before or after a candidate term, which we call *boundary words*. We follow the barrier word approach, where particular syntactic categories are used to delimit candidate terms, but develop this idea further by weighting boundary words according to their category. Analysis of boundary words occurring with terms and non-terms in the corpus showed that the category of the boundary word exerts an influence on its likelihood of being followed or preceded by a term. Table 1 depicts the frequency and percentage of boundary words in each of 4 syntactic categories accompanying terms and non-terms.

The weight for each category, shown in Table 2, is allocated according to its relative likelihood of occurring with a term as opposed to a non-term. A verb, therefore, occurring immediately before or after a candidate term, is statistically a better indicator of a term than an adjective is. By "a better indicator", we mean that a candidate term occurring with it is more likely to be valid. Each candidate term is assigned a syntactic weight, calculated by summing the category weights for the context boundary words occurring with it.

3.2 Terminological knowledge

Terminological knowledge concerns the terminological status of context words. A context word which is also a term (which we call a context term) is likely to be a better indicator than one which is not. The terminological status is determined by applying the NC-Value approach to the corpus, and considering the top third of the list of ranked results as valid terms. A context term (CT) weight is then produced for each candidate term, based on its total frequency of occurrence with all relevant context terms. The CT weight is formally described as follows:

$$CT(a) = \sum_{d \epsilon T_a} f_a(d) \tag{1}$$

where
a is the candidate term,
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.

3.3 Semantic knowledge

Semantic knowledge is obtained about context terms using the UMLS Metathesaurus and Semantic Network [29]. The former provides a semantic tag for each term, such as *Acquired Abnormality*. The latter provides a hierarchy of semantic types, from which we compute the similarity between a candidate term and the context terms it occurs with. An example of part of the network is shown in Figure 3.

Similarity is measured because we believe that a context term which is semantically similar to a candidate term is more likely to be significant than one which is less similar. We use the method for semantic distance described in [22], which is based on calculating the vertical position and horizontal distance between nodes in a hierarchy. Two weights are calculated:

- *positional*: measured by the combined distance from root to each node
- *commonality*: measured by the number of shared common ancestors multiplied by the number of words (usually two).

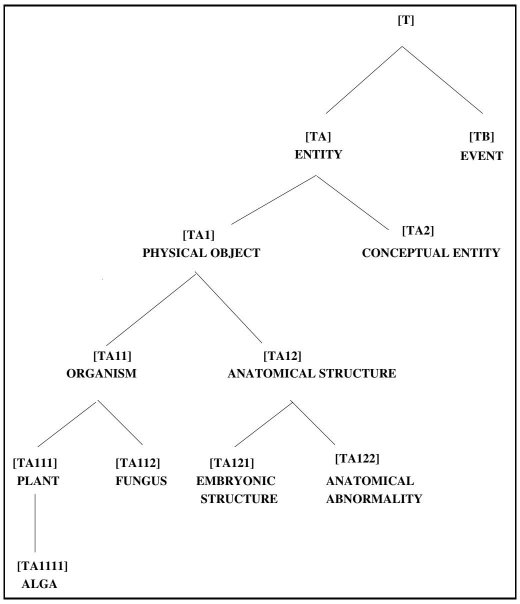

Figure 3: Fragment of the Semantic Network

Similarity between the 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. This is formally defined as follows:

$$\text{sim}(w_1...w_n) = \frac{\text{com}(w_1...w_n)}{\text{pos}(w_1...w_n)} \tag{2}$$

where
com($w_1...w_n$) is the commonality weight of words 1...n
pos($w_1...w_n$) is the positional weight of words 1...n.

Let us take an example from the UMLS. The similarity between a term belonging to the semantic category *Plant* and one belonging to the category *Fungus* would be calculated as follows:-

- *Plant* has the semantic code TA111 and *Fungus* has the semantic code TA112.
- The commonality weight is the number of nodes in common, multiplied by the number of terms we are considering. TA111

and TA112 have 4 nodes in common (T, TA, TA1 and TA11). So the weight will be $4 * 2 = 8$.

- The positional weight is the total height of each of the terms (where the root node has a height of 1). TA111 has a height of 5 (T, TA, TA1, TA11 and TA111), and TA112 also has a height of 5 (T, TA, TA1, TA11 and TA112). The weight will therefore be $5 + 5 = 10$.
- The similarity weight is the commonality weight divided by the positional weight, i.e. $8/10 = 0.8$.

4 Using Similarity for Term Sense Disambiguation

Although ambiguity is not as prolific in specialised texts as in general language, it can nevertheless cause problems for terminological applications. In the UMLS, there are many examples of multidimensionality. For example, the term *conjunctivitis* is classified in UMLS both as a *sign or symptom* and as a *disease or syndrome*. We shall refer to such a term as having two (or more) *meanings*, although the meanings may not strictly be different. (We could argue that the term has the same meaning in each case, but that there are different viewpoints from which it may be regarded).

There are essentially two issues at stake concerning term sense ambiguity. Firstly, it is beneficial to term extraction if we can classify the terms according to their meaning, so that if a term has two different meanings, we preserve this distinction in the extraction process. The importance of this depends on the ultimate use of the extracted terms. For applications such as information extraction, hypertext linking and machine translation, the granularity of distinction between term senses can vary considerably. For ontology creation or lexical tuning, it may be useful to make very fine-grained distinctions about meaning. Secondly, because of the way in which similarity is measured, if a term is positioned in two or more different places in the hierarchy, this creates problems, because a decision must be made as to which position should be used for the calculation.

The way in we measure similarity means that it is only really possible to deal with the second issue. This is because we do not examine each occurrence of a candidate term separately, but we look at all its occurrences in the text together. The similarity measure calculates similarity between the candidate term and every context term it occurs with, rather than just similarity between those context terms occurring with a particular instance of it. This means that we cannot use the similarity measure in this way to identify different meanings of a candidate term which occur in the same corpus. We can, however, use it to identify one particular meaning for that term within the corpus. This is more positive than it may appear, because experience shows that within a given corpus, different occurrences of a term are not so likely to have different meanings. Future work, however, could attempt to deal with this issue and extract terms according to their meaning in each instance.

4.1 Disambiguation of candidate terms

To disambiguate candidate terms using the similarity measure in this way, there are two intuitive possibilities. One solution would be to measure the similarity between the candidate term and context term for each meaning of the candidate term, and take an average of the results. However, this is liable to give a false impression of how relevant the context term is. For example, a context term might be highly relevant but might happen to have another meaning, situated in a distant part of the hierarchy, which is completely irrelevant. If the semantic distance is averaged out between these two, we would end up with a weighting far removed from either of the possibilities.

An alternative approach is to pick one of the two options, based on some kind of default criteria. This might be frequency-based or score-based. Since we use a relatively small corpus, the frequency-based option is not necessarily a wise choice, because each candidate term may only occur relatively infrequently. The score-based option, on the other hand, seems intuitive. This involves measuring the similarity between each meaning of the candidate term and the context terms it occurs with. The result with the highest score wins, because we assume that the more similar the context terms are to the candidate term, the more likely that that meaning is correct.

4.2 Some examples

To illustrate the disambiguation procedure, we shall give some examples from the corpus of eye pathology records which we use for our experiments.

4.2.1 Example 1

The candidate term *cholesterol_granuloma* has two possible semantic types associated with it:
1. Finding/Result (TA22)
2. Disease/Syndrome (TB222).
In the corpus, it occurs in the following contexts:

```
THE NODULE CONSISTS OF NUMEROUS FOREIGN BODY GIANT CELLS AND
HISTIOCYTES ARRANGED AROUND CHOLESTEROL CLEFTS, THE WHOLE
BEING EMBEDDED IN A FIBROUS STROMA. DIAGNOSIS: CHOLESTEROL GRANULOMA.
```

```
THE SUBRETINAL SPACE CONTAINS A MASS OF SEROUS EXUDATE WITH
SUBSTANTIAL HAEMORRHAGE. MANY PIGMENT LADEN MACROPHAGES ARE
PRESENT.  ONE SMALL CHOLESTEROL GRANULOMA  CAN BE SEEN POSTERIORLY.
```

From the first context, we can surmise that since the cholesterol granuloma is a *diagnosis*, the meaning *Finding/Result* is more relevant than the meaning *Disease/Syndrome*. Similarly, for the second context, the fact that the cholesterol granuloma can be *seen* also indicates that it is a *Finding*. The total similarity weight for meaning 1 is higher than for meaning 2 (2.0 compared with 0.6), which is supported by our intuition from the corpus. So in this case, the meaning *Finding/Result* would be correctly retained.

4.2.2 Example 2

The candidate term *degenerative_change* has the same two possible semantic types associated with it: *Finding/Result* and *Disease/Syndrome*. In the corpus, it occurs in contexts such as the following:

```
THE TISSUE CONSISTS OF A SMALL PIECE OF CONJUNCTIVAL TYPE
EPITHELIUM SHOWING DEGENERATIVE CHANGE IN THE COLLAGEN OF
THE SUBSTANTIA PROPRIA CONSISTENT WITH A PINGUECULA.
```

The total similarity score for the meaning *Finding/Result* is 20.727, while the score for the meaning *Disease/Syndrome* is 14.723. From the contexts, we consider that the first meaning is the best, and this is also the one with the highest score.

These examples by no means prove that the method is accurate. However, it is often very difficult to form a opinion about which meaning is most suited to a term in these kind of circumstances. For example, it is virtually impossible to say whether the meaning *Organ or Tissue Function* or the meaning *Disease or Syndrome* is more suitable for the term *graft_rejection*,in contexts such as:

```
THE PATHOLOGICAL FEATURES ARE CONSISTENT WITH LATE GRAFT
 REJECTION ALTHOUGH THERE IS NO EVIDENCE OF VASCULARISATION.
```

This is one of the reasons why an automatic method of disambiguation is so useful.

4.3 Disambiguation of context terms

The same method is also used when context terms are found to be ambiguous [21, 20]. In this case, similarity between candidate term and each meaning of that context term is measured, and again, the highest score wins. The meaning of the context term which has the lower score is simply discarded (for that particular candidate term). Since the meaning of the context term may be different in another situation in the text, the process is repeated when it is found occurring with another candidate term.

For example, in the corpus we find the candidate term *allergic conjunctivitis* occurring with the context term *conjunctivitis*. This context term has 2 possible positions in the hierarchy, as shown in Table 3. One of these belongs to the same node as the candidate term, producing a similarity of 1, whilst the other is in a completely different section of the hierarchy and only has a similarity of 0.4444.

A manual evaluation of a set of ambiguous context terms was carried out, to establish how well this approach worked. For each possible meaning of an ambiguous context term, we identified its semantic category (i.e. the label of its hierarchical position) and attempted to establish whether the meaning with the highest similarity score was in fact the one being used in that context. In some cases, it was not possible to differentiate between the two, since the meanings were very similar, but we estimated that in about 80% of

Candidate Term	Position	Context Term	Position	Similarity
allergic conjunctivitis	TB222	conjunctivitis	TA22	0.4444
allergic conjunctivitis	TB222	conjunctivitis	TB222	1
connective tissue	TA123	lying	TB11	0.1111
connective tissue	TA123	lying	TA125	0.8
epithelial cyst	TA12	lesions	TB222	0.1111
epithelial cyst	TA12	lesions	TA214	0.2222
epithelial cyst	TA12	lesions	TA22	0.4

Table 3: Similarity between candidate terms and ambiguous context terms

cases, the correct meaning was indeed the one with the highest similarity score. For example, the term *lying* can be classified either as a spatial concept or as a form of social behaviour. It occurs with the candidate term *connective tissue* in the context "*Lying among loose connective tissue is a collapsed cyst...*". Clearly the first meaning is the correct one in this case. The similarity between this meaning of *lying* and *connective tissue* is 0.8, whereas for the other meaning it is 0.1111, which corresponds with our intuitive interpretation.

For the ambiguous context terms, therefore, the scores are calculated for each meaning, and only the highest score (and associated meaning) is retained. If both meanings have an equal score, one meaning is picked at random, since for the time being we have no alternative method of selection.

5 The SNC-Value

The SNC-Value is calculated by combining the Importance Weight with the NC-Value, for each candidate term. Since the candidate terms have already been selected by the C-Value method, the SNC-Value does not attempt to select new terms or discard previously selected ones, but only to perform a re-ranking of those already chosen.

5.1 The NC-Value

The NC-Value comprises the bottom two layers of the model, and is calculated by combining the C-Value with a Context Factor.

5.1.1 The C-Value

The C-Value is defined as:

$$\text{C-Value}(a) = \left\{ \begin{array}{ll} log_2|a| \cdot f(a) & \text{a is not nested} \\ log_2|a| - \frac{1}{P(T_a)} \sum_{b \epsilon T_a} f(b)) & \text{is nested} \end{array} \right\}$$

It can be interpreted as follows. Two different cases apply: one for terms that are found as nested, and one for terms that are not. If a candidate string is not found as nested, its termhood is calculated from its total frequency and length. If it is found as nested, termhood is calculated from its total frequency, length, frequency as a nested string, and the number of longer candidate terms it appears in.

5.1.2 The Context Factor

The context factor is calculated for each term by summing the total weights for the context words occurring with it, multiplied by their frequency. The weight for a context word is based on the proportion of top-ranked terms occurring with it, defined as:

$$Weight(w) = \frac{f_w(t)}{n(t)} \tag{3}$$

where
w is the context word to be assigned a weight,
$Weight(w)$ is the weight assigned to word w,
t is a top-ranked term,
$f_w(t)$ is the number of terms occurring with word w,
$n(t)$ is the number of terms considered.

The context factor for a term is therefore defined as:

$$CF(t) = \sum_{w \epsilon C_t} f_t(w) \cdot weight(w) \tag{4}$$

where
t is the candidate term,
$C_t$is the set of context words of t,
w is a word from C_t,
$f_t(w)$ is the frequency of w as a context word of t,
$weight(w)$ is the weight of w as a context word,

5.2 The NC-Value

Combining the C-Value with the Context Factor, we get the NC-Value, as follows:

$$NCvalue(a) = 0.8 * Cvalue(a) + 0.2 * CF(a) \tag{5}$$

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.

The factors 0.8 and 0.2 were determined through a set of empirical experiments to find the best ratio.

5.3 The Importance Weight

The three weights described in Section 3 (syntactic, terminological and semantic) are calculated for all relevant context words or context terms. The total weights for the context are then combined according to the following equation:

$$IW(a) = \sum_{b \epsilon C_a} syn_a(b) + \sum_{d \epsilon T_a} f_a(d) \cdot sim_a(d) \tag{6}$$

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,
$syn_a(b)$ is the syntactic weight of b as a context word of a,
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,
$sim_a(d)$ is the similarity weight of d as a context term of a.

This states that the Importance Weight is composed of the total terminological weight, multiplied by the total semantic weight, and then added to the total syntactic weight of all the context words or context terms related to the candidate term.

5.4 The SNC-Value

The Importance Weight gives a score for each candidate term based on the importance of the contextual information surrounding it. To obtain the final SNC-Value ranking, the Information Weight is combined with the NC-Value, as expressed formally below:

$$SNCValue(a) = NCValue(a) + IW(a) \tag{7}$$

where
a is the candidate term
$NCValue(a)$ is the NC-Value of a
IW is the Importance Weight of a.

6 Evaluation

The SNC-Value method was initially tested on a small corpus of eye pathology records, which had been tagged with the Brill part-of-speech tagger [5]. The candidate terms were first extracted using the NC-Value method [13], and the SNC-Value was then calculated. To evaluate the results, we examined both the performance of the similarity weight, and the overall performance of the system.

An example of the final result is shown in Table 4. This compares the top 20 results from the SNC-Value list with the top 20 from the NC-Value list. The terms in italics are those which were considered as not valid. We can note here two points. Firstly, the weights for the SNC-Value are substantially greater than those for the NC-Value. This, in itself, is not important, since it is the position in the list, i.e. the *relative* weight, rather than the *absolute* weight, which is important. Secondly, we can see that there are more valid terms in the SNC-Value results than in the NC-Value results. It is hard to make further judgements based on this list alone, because we cannot say whether one term is better than another, if the two terms are both valid.

6.1 Evaluation methods

The main evaluation procedure was carried out with respect to a manual assessment of the list of terms by 2 domain experts. There

Term	SNC	Term	NC
bowman's_membrane	605782	*plane_of_section*	1752.71
malignant_melanoma	231237	descemet's_membrane	1345.76
hyaline_fibrous_tissue	215843	basal_cell_carcinoma	1268.21
planes_of_section	170016	*stump_of_optic_nerve*	993.15
trabecular_meshwork	157353	basal_cell_papilloma	616.614
keratinous_debris	101644	*plane_of_section=*	506.517
bruch's_membrane	94996.2	*melanoma_of_choroid*	497.673
plane_of_section=	90109.4	*planes_of_section*	453.716
melanoma_of_choroid	71615.1	malignant_melanoma	448.591
lymphocytic_infiltration	53822	optic_nerve_head	422.211
ciliary_processes	52355.7	ciliary_processes	421.204
cellular_fibrous_tissue	51486.8	bruch's_membrane	413.027
squamous_epithelium	46928.9	keratinous_cyst	392.944
optic_nerve_head	39054.5	*ellipse_of_skin*	267.636
pupillary_border	36510.8	*wedge_of_lid_margin*	211.414
corneal_epithelium	31335.9	*scar_track*	228.217
scleral_invasion	31017.4	connective_tissue	167.053
granulation_tissue	28010.1	*vertical_plane*	167.015
stratified_squamous_epithelium	27445.5	carcinoma_of_lid	164
ocular_structures	26143.6	*excision_biopsy*	155.257

Table 4: Top 20 results for the SNC-Value and NC-Value

are, however, problems associated with such an evaluation. Firstly, there is no gold standard of evaluation, and secondly, manual evaluation is both fallible and subjective. To avoid this problem, we measure the performance of the system in relative terms rather than in absolute terms, by measuring the improvement over the results of the NC-Value as compared with manual evaluation. Although we could have used the list of terms provided in the UMLS, instead of a manually evaluated list, we found that there was a huge discrepancy between this list and the list validated by the manual experts (only 20% of the terms they judged valid were found in the UMLS). There are also further limitations to the UMLS, such as the fact that it is only specific to medicine in general, but not to eye pathology, and the fact that it is organised in such a way that only the preferred terms, and not lexical variants, are actively and consistently present.

We first evaluate the similarity weight individually, since this is the main principle on which the SNC-Value method relies. We then evaluate the SNC-Value as a whole by comparing it with the NC-Value, so that we can evaluate the impact of the addition of the deeper forms of linguistic information incorporated in the Importance Weight.

6.2 Similarity Weight

One of the problems with our method of calculating similarity is that it relies on a pre-existing lexical resource, which means it is prone to errors and omissions. Bearing in mind its innate inadequacies, we can nevertheless evaluate the *expected theoretical performance* of the measure by concerning ourselves only with what is covered by the thesaurus. This means that we assume completeness (although we know that this is not the case) and evaluate it accordingly, ignoring anything which may be missing.

The semantic weight is based on the premise that the more similar a context term is to the candidate term it occurs with, the better an indicator that context term is. So the higher the total semantic weight for the candidate term, the higher the ranking of the term and the better the chance that the candidate term is a valid one. To test the performance of the semantic weight, we sorted the terms in descending order of their semantic weights and divided the list into 3, such that the top third contained the terms with the highest

Section	Term	Non-Term
top set	76%	24%
middle set	56%	44%
bottom set	49%	51%

Table 5: Semantic weights of terms and non-terms

semantic weights, and the bottom third contained those with the lowest. We then compared how many valid and non-valid terms (according to the manual evaluation) were contained in each section of the list.

The results, depicted in Table 5, can be interpreted as follows. In the top third of the list, 76% were terms and 24% were non-terms, whilst in the middle third, 56% were terms and 44% were non-terms, and so on. This means that most of the valid terms are contained in the top third of the list and the fewest valid terms are contained in the bottom third of the list. Also, the proportion of terms to non-terms in the top of the list is such that there are more terms than non-terms, whereas in the bottom of the list there are more non-terms than terms. This therefore demonstrates two things:

- more of the terms with the highest semantic weights are valid, and fewer of those with the lowest semantic weights are valid;
- more valid terms have high semantic weights than non-terms, and more non-terms have lower semantic weights than valid terms.

We also tested the similarity measure to see whether adding some statistical information would improve its results, and regulate any discrepancies in the uniformity of the hierarchy. The methods which intuitively seem most plausible are based on information content. e.g.[30, 34]. The information content of a node is related to its probability of occurrence in the corpus. The more frequently it appears, the more likely it is to be important in terms of conveying information, and therefore the higher weighting it should receive. We performed experiments to compare two such methods with our similarity measure. The first considers the probability of the MSCA of the two terms (the lowest node which is an ancestor of both), whilst the second considers the probability of the nodes of the terms being compared. However, the findings showed a negligible difference

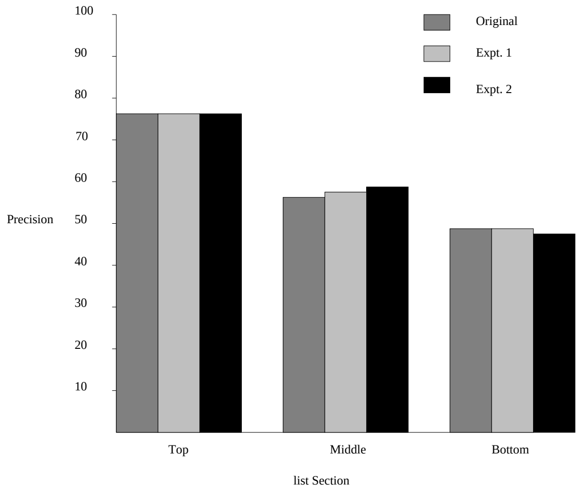

Figure 4: Influence of statistical methods on the similarity measurement

between the three methods, as depicted in Table 4, so we conclude that there is no advantage to be gained by adding statistical information, for this particular corpus. It is possible that with a larger corpus or different hierarchy, this might not be the case.

6.3 Overall Evaluation of the SNC-Value

We first compare the precision rates for the SNC-Value and the NC-Value (Table 6), by dividing the ranked lists into 10 equal sections. Each section contains 250 terms, marked as valid or invalid by the manual experts. In the top section, the precision is higher for the SNC-Value, and in the bottom section, it is lower. This indicates that the precision span is greater for the SNC-Value, and therefore that the ranking is improved. The distribution of valid terms is also better for the SNC-Value, since of the valid terms, more appear at the top of the list than at the bottom. The table also shows the error rates between the two methods, for each section of the list. From this it is apparent that while there is not a marked distinction at the very top of the table, further down the list, there is much more discrepancy between the two results. For example, in the 9th section there is an error rate of 41%, which is indeed a marked difference. it should be pointed out, however, that the error is not necessarily in favour of the SNC-Value at all points.

	SNC-Value		NC-Value		
Section	Valid	Precision	Valid	Precision	Error Rate
1	163	64%	160	62%	3%
2	84	33%	98	38%	13%
3	89	35%	69	27%	23%
4	89	35%	78	30%	14%
5	76	30%	87	34%	12%
6	57	22%	78	30%	27%
7	66	26%	92	36%	28%
8	75	29%	100	39%	26%
9	70	27%	42	16%	41%
10	59	23%	68	27%	15%

Table 6: Precision of SNC-Value and NC-Value

These discrepancies are also indicated in Figure 5. Here we can see that the SNC-Value graph is smoother than that of the NC-Value. We can compare the graphs more accurately using a method we call *comparative upward trend.* Because there is no one ideal graph, we instead measure how much each graph deviates from a monotonic line downwards. This is calculated by dividing the total rise in precision percentage by the length of the graph. A graph with a lower upward trend will therefore be better than a graph with a higher upward trend. If we compare the upward trends of the two graphs, we find that the trend for the SNC-Value is 0.9, whereas the trend for the NC-Value is 2.7. This again shows that the SNC-Value ranking is better than the NC-Value ranking, since it is more consistent.

Table 7 shows a more precise investigation of the top portion of the list (where it is to be expected that terms are most likely to be valid, and which is therefore the most important part of the list) shows that the precision is most improved here, both in terms of accuracy and in terms of distribution of weights. At the bottom of the top section, the precision is much higher for the SNC-Value. This is important because ideally, all the terms in this part of the list should be valid,

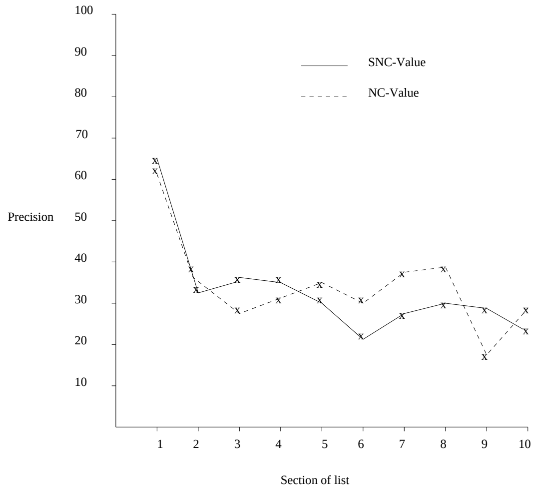

Figure 5: Precision of SNC-Value and NC-Value

	SNC-Value		NC-Value	
Section	Valid	Precision	Valid	Precision
1	21	84%	19	76%
2	19	76%	23	92%
3	17	68%	21	84%
4	16	64%	13	52%
5	18	72%	13	52%
6	12	48%	19	76%
7	13	52%	18	72%
8	17	68%	14	56%
9	13	52%	10	40%
10	14	56%	8	32%

Table 7: Precision of SNC-Value and NC-Value for top 250 terms

7 Conclusions

In this paper, we have described a method for multi-word term extraction which improves on traditional statistical approaches by incorporating more specific contextual information. It focuses particularly on measuring the strength of association (in semantic terms) between a candidate term and its context. Evaluation shows improvement over the NC-Value approach. Although the improvement in precision overall appears small (largely because we have used a very small corpus for testing), there are other factors to be considered. In particular, it should be noted that the distribution and quality of terms in the list is much improved.

The contextual information acquired can also be used for a number of other related tasks, such as disambiguation and clustering. At present, the semantic information is acquired from a pre-existing domain-specific thesaurus, but there are possibilities for creating such a thesaurus automatically, or enhancing an existing one, using the contextual information we acquire [35, 23].

There is much scope for further extensions of this research. Firstly, it could be extended to other domains and larger corpora, in order to see the true benefit of such an approach. Secondly, the thesaurus could be tailored to the corpus, as we have mentioned. An incremental approach might be possible, whereby the similarity measure is combined with statistical information to tune an existing ontology. Also, the UMLS is not designed as a linguistic resource, but as an information resource. Some kind of integration of the two types of resource would be useful so that, for example, lexical variation could be more easily handled.

References

[1] S. Ananiadou. A methodology for automatic term recognition. In *Proc. of 15th International Conference on Computational Linguistics (COLING)*, Kyoto, Japan, 1994.

[2] M. S. Blois. Medicine and the nature of vertical reasoning. *New England Journal of Medicine*, 318(13):847–851, 1988.

[3] D. Bourigault. Surface grammatical analysis for the extraction of terminological noun phrases. In *Proc. of 14th Internation-*

al Conference on Computational Linguistics (COLING), pages 977–981, Nantes, France, 1992.

[4] Catherine Bowman, Diane Michaud, and Heidi Suonuuti. Do's and don'ts of terminology management. In S.E. Wright and G. Budin, editors, *Handbook of Terminology Management*, volume 1: Basic Aspects of Terminology Management, chapter 2.2.4, pages 215–217. John Benjamins, Amsterdam, 1997.

[5] Eric Brill. A simple rule-based part of speech tagger. In *Proc. of 3rd Conference of Applied Natural Language Processing*, 1992.

[6] Ido Dagan, Fernando Pereira, and Lillian Lee. Similarity-based estimation of word co-occurrence probabilities. In *Proc. of the 32nd Annual Meeting of the Association for COmputational Linguistics (ACL'94)*, pages 272–278, 1994.

[7] B. Daille, E. Gaussier, and J.M. Langé. Towards automatic extraction of monolingual and bilingual terminology. In *Proc. of 15th International Conference on Computational Linguistics (COLING)*, pages 515–521, 1994.

[8] L. Davidson, J. Kavanagh, K. Mackintosh, I. Meyer, and D. Skuce. Semi-automatic extraction of knowledge-rich contexts from corpora. In *Proc. of 1st Workshop on Computational Terminology (Computerm '98)*, pages 50–56, Montreal, Canada, 1998. COLING-ACL.

[9] P.F. De Vries Robbé and F.J. Flier. Towards a common vocabulary for classification and definition. *Terminology*, 1(1):97–102, 1994.

[10] Robert Dubuc and Andy Lauriston. Terms and contexts. In S.E. Wright and G. Budin, editors, *Handbook of Terminology Management*, volume 1: Basic Aspects of Terminology Management, chapter 1.3.3, pages 80–87. John Benjamins, Amsterdam, 1997.

[11] Chantal Enguehard and Laurent Pantera. Automatic natural acquisition of a terminology. *Journal of Quantitative Linguistics*, 2(1):27–32, 1994.

[12] David A. Evans and Chengxiang Zhai. Noun phrase analysis in unrestricted text for information retrieval. In *Proc. of 34th Annual Meeting of the Assocation for Computational Linguistics (ACL'96)*, pages 17–24, Santa Cruz, USA, 1996.

[13] K.T. Frantzi. *Automatic Recognition of Multi-Word Terms*. PhD thesis, Manchester Metropolitan University, England, 1998.

[14] K.T. Frantzi and S. Ananiadou. Extracting nested collocations. In *Proc. of the 16th International Conference on Computational Linguistics (COLING)*, pages 41–46, Copenhagen, Denmark, 1996.

[15] K.T. Frantzi and S. Ananiadou. The C-Value/NC-Value domain independent method for multi-word term extraction. *Journal of Natural Language Processing*, 6(3):145–179, 1999.

[16] G. Grefenstette. *Explorations in Automatic Thesaurus Discovery*. Kluwer Academic Publishers, 1994.

[17] ISO-704. International organisation for standardisation (ISO) - principles and methods of terminology, 1987.

[18] J.S. Justeson and S.M. Katz. Technical terminology: some linguistic properties and an algorithm for identification in text. *Natural Language Engineering*, 1:9–27, 1995.

[19] Andy Lauriston. *Automatic term recognition: performance of linguistic and statistical learning techniques*. PhD thesis, UMIST, Manchester, UK, 1996.

[20] D.G. Maynard and S. Ananiadou. Acquiring contextual information for term disambiguation. In *Proc. of 1st Workshop on Computational Terminology, Computerm '98*, pages 86–90, Montreal, Canada, 1998.

[21] D.G. Maynard and S. Ananiadou. Term sense disambiguation using a domain-specific thesaurus. In *Proc. of 1st International Conference on Language Resources and Evaluation (LREC)*, pages 681–687, Granada, Spain, 1998.

[22] D.G. Maynard and S. Ananiadou. Identifying contextual information for term extraction. In *Proc. of 5th International*

Congress on Terminology and Knowledge Engineering (TKE '99), pages 212–221, Innsbruck, Austria, 1999.

[23] D.G. Maynard and S. Ananiadou. A linguistic approach to context clustering. In *Proc.of Natural Language Processing Pacific Rim Symposium (NLPRS)*, pages 346–351, Beijing, China, 1999.

[24] D.G. Maynard and S. Ananiadou. Term extraction using a similarity-based approach. In *Recent Advances in Computational Terminology*. John Benjamins, 1999. to appear.

[25] Merriam-Webster. Hypertext Webster Dictionary, 1999. http://www.fin.gov.nt.ca/webster.htm.

[26] I. Meyer, K. Mackintosh, C. Barrière, and T. Morgan. Conceptual sampling for terminographical corpus analysis. In *Proc. of 5th International Congress on Terminology and Knowledge Engineering (TKE '99)*, pages 256–267, Innsbruck, Austria, 1999.

[27] Hiroshi Nakagawa and Tatsunori Mori. Nested collocation and compound noun for term extraction. In *Proc. of 1st Workshop on Computational Terminology, Computerm '98*, pages 64–70, Montreal, Canada, 1998.

[28] S.J. Nelson, N.E. Olson, L. Fuller, M.S. Tuttle, W.G. Cole, and D.D. Sherertz. Identifying concepts in medical knowledge. In *Proc. of 8th World Congress on Medical Informatics (MEDINFO)*, pages 33–36, 1995.

[29] NLM. *UMLS Knowledge Sources*. National Library of Medicine, U.S. Dept. of Health and Human Services, 8th edition, January 1997.

[30] P. Resnik. Disambiguating noun groupings with respect to WordNet senses. In *Proc. of 3rd Workshop on Very Large Corpora*. MIT, 1995.

[31] Guy Rondeau. Introduction to terminology. In Juan C. Sager, editor, *Readings in Terminology*. CCL, UMIST, 1987. (English version of Chapters 1-5, edited by J.C. Sager 1980).

[32] J.C. Sager. *A Practical Course in Terminology Processing*. John Benjamins, 1990.

[33] Frank Smadja. Retrieving collocations from text: Xtract. *Computational Linguistics*, 19(1):143–177, 1993.

[34] A. Smeaton and I. Quigley. 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, 1996.

[35] Akira Ushioda. Hierarchical clustering of words. In *Proc. of 16th International Conference on Computational Linguistics (COLING)*, pages 1159–1162, 1996.