Identifying Terms by their Family and Friends

Diana Maynard
Dept. of Computer Science
University of Sheffield
Regent Court, 211 Portobello St
Sheffield, S1 4DP, UK
d.maynard@dcs.shef.ac.uk

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

Abstract

Multi-word terms are traditionally identified using statistical techniques or, more recently, using hybrid techniques combining statistics with shallow linguistic information. 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. 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

Although statistical approaches to automatic term recognition, e.g. (Bourigault, 1992; Daille et al., 1994; Enguehard and Pantera, 1994; Justeson and Katz, 1995; Lauriston, 1996), 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. Although contextual information has been previously used, e.g. in general language (Grefenstette, 1994) and in the NC-Value method for term recognition (Frantzi, 1998; Frantzi and Ananiadou, 1999), only shallow syntactic information is used in these cases. The TRUCKS approach identifies different elements of the context which are combined to form the Information Weight, a measure of how strongly related the context is to a candidate term. The Information Weight is then combined with the statistical information about a candidate term and its context, acquired using the NC-Value method, to form the SNC-Value. Section 2 describes the NC-Value method. Section 3 discusses the importance of contextual information and explains how this is acquired. Sections 4 and 5 describe the Information Weight and the SNC-Value respectively. We finish with an evaluation of the method and draw some conclusions about the work and its future.

2 The NC-Value method

The NC-Value method uses a combination of linguistic and statistical information. Terms are first extracted from a corpus using the C-Value method (Frantzi and Ananiadou, 1999), a measure based on frequency of occurrence and term length. This is defined formally as:

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

where
a is the candidate string,
$f(a)$ is its frequency in the corpus,
ϵT_a is the set of candidate terms that contain a,
$P(T_a)$ is the number of these candidate terms.

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.

The NC-Value method builds on this by incorporating contextual information in the form of a context factor for each candidate term. A context word can be any noun, adjective or verb appearing within a fixed-size window of the candidate term. Each context word is assigned a weight, based on how frequently it appears with a candidate term. These weights are then summed for all context words relative to a candidate term. The Context Factor is combined with the C-Value to form the NC-Value:

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

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.

3 Contextual Information: a Term's Social Life

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. We acquire this knowledge by extracting three different types of contextual information:

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

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*. Following "barrier word" approaches to term recognition (Bourigault, 1992; Nelson et al., 1995), where particular syntactic categories are used to delimit candidate terms, we develop this idea further by weighting boundary words according to their category. The weight for each category, shown in Table 1, 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.

Category	Weight
Verb	1.2
Prep	1.1
Noun	0.9
Adj	0.7

Table 1: Weights for categories of boundary words

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:

$$CT(a) = \sum_{d \epsilon T_a} f_a(d) \quad (2)$$

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 (NLM, 1997). 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 1.

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 (Maynard and Ananiadou, 1999a), 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).

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

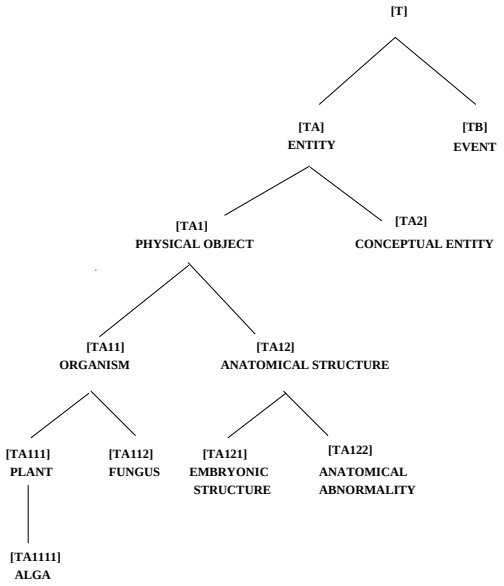

Figure 1: Fragment of the Semantic Network

where the two nodes are identical, and 0 being the case where there is no common ancestor. This is formally defined as follows:

$$\mathrm{sim}(w_1 ... w_n) = \frac{\mathrm{com}(w_1 ... w_n)}{\mathrm{pos}(w_1 ... w_n)} \qquad (3)$$

where

$\mathrm{com}(w_1 ... w_n)$ is the commonality weight of words 1...n

$\mathrm{pos}(w_1 ... w_n)$ 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 The Information Weight

The three individual weights described above are calculated for all relevant context words or context terms. The total weights for the context are then combined according to the following equation:

$$IW(a) = \sum_{b \epsilon C_a} syn_a(b) + \sum_{d \epsilon T_a} f_a(d) \cdot sim_a(d) \qquad (4)$$

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 basically means that the Information 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 The SNC-Value

The Information 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 statistical information obtained using the NC-Value method, as expressed formally below:

$$SNCValue(a) = NCValue(a) + IW(a) \qquad (5)$$

where

a is the candidate term

$NCValue(a)$ is the NC-Value of a

IW is the Importance Weight of a

For details of the NC-Value, see (Frantzi and Ananiadou, 1999).

An example of the final result is shown in Table 2. 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 shall discuss the results in more detail in the next section, but we can note here three 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

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 2: Top 20 results for the SNC-Value and NC-Value

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. Thirdly, we can see that more of the top 20 terms are valid for the SNC-Value than for the NC-Value: 17 (85%) as opposed to 10 (50%).

6 Evaluation

The SNC-Value method was initially tested on a corpus of 800,000 eye pathology reports, which had been tagged with the Brill part-of-speech tagger (Brill, 1992). The candidate terms were first extracted using the NC-Value method (Frantzi, 1998), and the SNC-Value was then calculated. To evaluate the results, we examined the performance of the similarity weight alone, and the overall performance of the system.

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 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

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

Table 3: Semantic weights of terms and non-terms

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 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 3, 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.(Resnik, 1995; Smeaton and Quigley, 1996). 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 between the three methods, 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.

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

Table 4: Precision of SNC-Value and NC-Value

6.3 Overall Evaluation of the SNC-Value

We first compare the precision rates for the SNC-Value and the NC-Value (Table 4), 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.

Looking at Figure 2, 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 5 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) We see 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

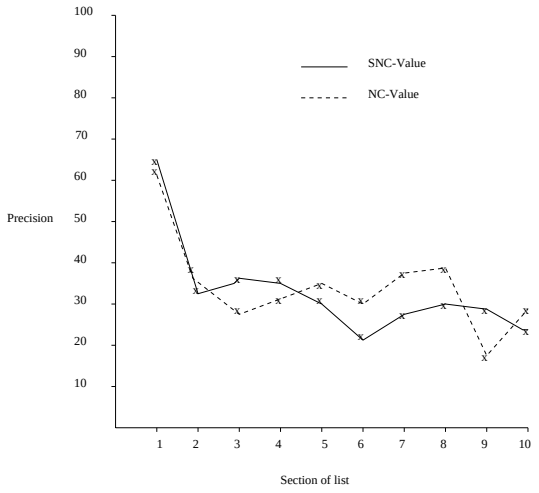

Figure 2: Precision of SNC-Value and NC-Value

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

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,

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 percentages are small. This is largely because we have used a very small corpus for testing.

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 (Ushioda, 1996; Maynard and Ananiadou, 1999b).

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

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

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

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

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

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

K.T. Frantzi. 1998. *Automatic Recognition of Multi-Word Terms.* Ph.D. thesis, Manchester Metropolitan University, England.

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

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

Andy Lauriston. 1996. *Automatic term recognition: performance of linguistic and statistical learning techniques.* Ph.D. thesis, UMIST, Manchester, UK.

D.G. Maynard and S. Ananiadou. 1999a. Identifying contextual information for term extraction. In *Proc. of 5th International Congress on Terminol-*

ogy and Knowledge Engineering (TKE '99), pages 212–221, Innsbruck, Austria.

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

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

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

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

A. Smeaton and I. Quigley. 1996. Experiments on using semantic distances between words in image caption retrieval. In *Proc. of 19th International Conference on Research and Development in Information Retrieval*, Zurich, Switzerland.

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