Incorporating Linguistic Information for Multi-Word Term Extraction

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

January 9, 1999

Abstract

Methods for multi-word term extraction generally involve statistical and/or linguistic techniques, but the linguistic information used is mainly in the form of simple syntactic information. Our approach makes use of semantic information from UMLS [12], a domain-specific thesaurus, and of a rich syntactic and semantic representation. by incorporation new contextual weights based on these, and in particular, a similarity measure we have developed, we improve on the NC-Value method of term recognition [6], which produces a ranking of candidate terms. Our results show that by adding deeper forms of information to this value, we can make more of a distinction between terms and non-terms, thereby improving the ranking, and we can also perform some disambiguation of these terms.

1 Introduction

Whilst the use of contextual information is clearly beneficial for multi-word terms, it is difficult to identify which parts of the context are most useful and what type of information is relevant. The context is generally considered simply as a "bag of words" and its structure and relationship to the term in question are largely ignored.

Our approach makes use of both lexical resources and corpus-based information. Semantic information is acquired from UMLS [12], a domain-specific thesaurus, and syntactic information is extracted from the corpus by means of analysers producing a rich linguistic representation. We improve on a measure called NC-Value developed in [7], which produces a ranking of candidate terms. Our results show that by adding deeper forms of information to this value, we can improve this ranking, thereby making more of a distinction between terms and non-terms.

With respect to the problem of identifying relevant parts of the context, we apply several measures. We consider the terminological status of each context word, its frequency, and semantic and syntactic information about both individual words and chunks of context. This is used to cluster the contexts according to their properties, so that we can identify similar contexts for use as term indicators or disambiguators.

2 Related Work

Methods for multi-word term extraction tend to be linguistic [2], [1] or hybrid (linguistic and statistical) e.g. [4],[9],[6]. However, the linguistic information used is generally very shallow and consists mainly of simple syntactic information in the form of a grammar restricting the types of phrases to be extracted as potential terms. We believe that a deeper understanding of the conceptual aspect of terms is necessary [3], and to this extent we propose the incorporation of semantic information for both term disambiguation [10] and extraction. The similarity measure we propose is based on methods found more commonly in machine translation, e.g. [16], [14]. These generally combine some kind of hierarchical distance weight with an importance weight based on frequency of occurrence. Unlike other context similarity weights, e.g. [15], [13], our weight is not based on statistical but semantic properties of the context terms.

3 Extracting Contextual Information

The NC value method uses linguistic information from the context to help identify terms. However, Frantzi et al. only make use of limited linguistic information, in the form of a simple syntactic filter which identifies nouns, verbs and adjectives in the context. It does not make any distinction between these categories, nor does it take into account either the relative position of the context word or its semantic properties [5].

We extend this idea by considering the status of the context word, and the syntactic and semantic relationships involved. We consider 4 main issues:

1. A context word which is a term is more likely to be significant as a term indicator than a context word which is a non-term.

2. Of these context terms, we propose that those which have similar semantic properties to the candidate term are better indicators.

3. The position of relevant context words with respect to the term may be important.

4. The syntactic category of the context word and possibly its syntactic relationship with the term should be considered. Some categories intuitively convey more useful information than others.

Our measure combines these factors, in the form of weights assigned to the candidate term, as outlined below.

3.1 Context Term Weight

Since we do not know in advance which context words are terms, this step can only be undertaken once we have a preliminary list of candidate terms. For this we use the top of the list of terms extracted by the C-value approach [6], since this should contain the "best" terms (or, at least, those which behave in the most term-like fashion). A context term (CT) weight is assigned to each candidate term based on how frequently it appears with a context term.

3.2 Similarity Weight

Similarity is calculated using a thesaurus hierarchy, based on the approach of [14]. For this we use UMLS, a domain-specific thesaurus, to assign semantic categories to each word in question. The similarity between a context term and a candidate term is then measured by calculating the distance between their semantic categories in the hierarchy provided by the UMLS Semantic Network. The semantic distance is defined in terms of a positional weight (the vertical position of the Most Specific Common Abstraction of the two nodes), and a commonality weight (the number of shared common ancestors of the two nodes), as follows:

$$\mathrm{sim}(w_1 \ldots w_n) = \frac{n(\mathrm{com}(w_1 \ldots w_n))}{\mathrm{pos}(w_1 \ldots w_n)}$$

where

n is the number of words being compared;
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.

The similarity weight is also used to separate the meanings of ambiguous terms. If either the candidate term or the context term is ambiguous, it will usually have more than one semantic category, and each meaning of the term will thus receive a different similarity weight. For a more detailed description of these weights, see [11], [10].

3.3 Syntactic Weight

This phase is still currently under development, but involves using a rich syntactic and semantic representation provided by analysers developed at the University of Pennsylvania [8] to provide detailed linguistic information about the context words. The corpus is analysedusing a syntactic chunker, a dependency analyser and a tagger. This provides us with syntactic categories and dependency relations between "chunks" of text. This information is then used to depict the relations between terms and context, which can be used to help identify important contexts. The weight is primarily dependent on the position of the context word (preceding or following the term) and its syntactic category (since some categories provide more useful information than others.

Term	Ranking	NC-value	Similarity Weight	CT Weight
plane of section	1752.71	1107.11	27.5429	645.6
descemet's membrane	1671.6	1328.8	17.8571	342.8
optic nerve	1670.75	1652.75	162.114	230.2
basal cell carcinoma	1485.31	1257.39	108.407	216.6
fibrous tissue	1359.73	1121.73	158.15	238
anterior chamber	1322.34	1015.74	236.4	306.6
bowman's membrane	1154.09	863.693	10.4643	290.4
corneal diameters	947.777	874.577	4.88571	73.2
cell carcinoma	892.491	760.691	55.7857	131.8
malignant melanoma	715.904	589.389	57.19	117.8

Table 1: Similarity and Context Term Weights for the 10 top-ranked terms

4 Improving the NC-Value Ranking

By incorporating the first two weights into our measure, we have been able to both improve the ranking of candidate terms from the NC-value, and distinguish between certain cases of term ambiguity. This means that there is more of a distinction drawn between terms and non-terms, with most valid terms receiving a higher score and thus moving further up the ranking. Table 1 shows the 10 top-ranked terms and their new ranking, along with their old NC-value ranking and the two weights we have introduced - the Similarity and Context Term Weights.

5 Conclusions and Further Work

This paper has focused on the use of semantic information to improve multi-word term extraction. Making use of

This paper has focused on the use of semantic information to improve multi-word term extraction. Making use of contextual information is not sufficient unless we have some means of distinguishing the relative contribution that context words can make. Our aim is for a deeper understanding of the nature of terms and their relationship with the concepts they represent, which should contribute not only towards methods of automatic extraction and disambiguation, but also towards theoretical knowledge of terminology.

Work on developing the syntactic weights is currently under development. We are evaluating how best to weight the different syntactic categories and position of the context words. We are also investigating the creation of clusters of contexts, based on syntactic and semantic properties. By normalising the syntactic structure of the contexts and then creating semantic frames, we are able to create clusters of contexts with the same deep syntactic structure and similar semantic properties. It remains to be seen exactly how this information can best be used to aid our method of term extraction.

6 Acknowledgements

We would like to thank Aravind Joshi and B. Srinivas at the University of Pennsylvania for allowing us to make use of their Supertagger and other analysers.

References

[1] S. Ananiadou. A methodology for automatic term recognition. In Proc. of COLING, Kyoto, Japan, 1994.

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

[3] L. Bowker. A multidimensional approach to classification in Terminology: Working with a computational framework. PhD thesis, University of Manchester, England, 1995.

[4] B. Daille, E. Gaussier, and J.M. Lang'e. Towards automatic extraction of monolingual and bilingual terminology. In Proc. of COLING 94, pages 515-521, 1994.

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

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

[7] K.T. Frantzi, S. Ananiadou, and J. Tsujii. The c-value/nc-value method of automatic recognition for multi-word terms. In Lecture Notes in Computer Science, volume 1513, pages 585-600. Springer-Verlag, 1998.

[8] A.K. Joshi and B Srinivas. Disambiguation of super parts of speech (or supertags): Almost parsing. In Proc. of COLING-94, Kyoto, Japan, 1994.

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

[10] D. Maynard and S. Ananiadou. Identifying contextual information for term extraction. In Proc. of 1st Workshop on Computational Terminology, Computerm '98, Montreal, Canada, 1998.

[11] D. Maynard and S. Ananiadou. Term sense disambiguation using a domain specific thesaurus. In Proc. of 1st International Conference on Language Resources and Evaluation, Granada, Spain, 1998.

[12] NLM, U.S. Dept. of Health and Human Services. UMLS Knowledge Sources, 8th edition, January 1997.

[13] H. Schutze. Dimensions of meaning. In Proc. of Supercomputing '92, 1992.

[14] E. Sumita and H. Iida. 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, 1991.

[15] Chengziang Zhai. Exploiting context to identify lexical atoms - a statistical view of linguistic context. In proc. of International & Interdisciplinary Conference on Modifying and Using Context (CONTEXT-97), pages 119- 129, Rio de Janeiro, Brazil, 1997.

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