Comparative Evaluation of C-value in the Treatment of Nested Terms

Abstract

For term extraction purposes, terms are essentially multi-word units with the additional property of "termhood". While nesting, i.e. the occurrence of smaller units within a larger lexical unit, is usually not the primary concern in collocation extraction for general language applications, it presents a core problem in term extraction systems. A term extractor should be able to, firstly, identify an instance of nesting and, secondly, decide which of the nested candidates are to be extracted.

In robust applications that involve little or no linguistic pre-processing, adequate treatment of nesting may be essential to the overall performance of the term extractor. If, for example, the trigram *axial compressive force* is extracted from an untagged corpus, the system will most probably detect also the statistical relevance of the bigrams *axial compressive* and *compressive force*. While the latter may indeed be terminologically relevant, the former is clearly noise and does not correspond to any of the typical terminological patterns in English.

Even in systems which use morphosyntactic analysis the ranking of nested terms may not be an easy task. Consider for example the collocation *reactor coolant system replacement* and the nested bi- and trigrams it contains. Should *reactor coolant* and *coolant system* both be retained, while *system replacement* rejected? Is the extracted fourgram at all terminologically relevant? Clearly such decisions depend on the purpose of the term extraction task and on the requirements of the target users.

Several authors have proposed methods for the treatment of nested terms. A well-known but not very widely applied method is C-value (Frantzi/Ananiadou 1996, Nakagawa/Mori 1998), which ranks terms according to their stability in the corpus. C-value is defined as:

$$C-value(a) = (length(a)-1)(\,freq(a) - \frac{t(a)}{c(a)})$$

where “a” is a collocation, t(a) is the frequency of “a” in longer candidates of collocations and c(a) is the number of longer candidates of collocations including “a”.

An alternative method proposed by Silva et al. (1999) uses LocalMax, an algorithm that measures the "glue" between words and proposes possible term boundaries. It is, however, our observation that many term extraction methods involve no explicit strategy to tackle this problem, which probably means that in cases of nesting no additional ranking is performed and all variants are extracted if they fulfil other criteria of *termhood*.

We report on an implementation of C-value for the treatment of nested terms in a bilingual term extraction system. The system was designed for English and Slovene in two parallel versions: the statistical version works with raw texts and no morphosyntactic tagging to extract term candidates monolingually and then identify translation equivalents, while the hybrid version uses tagged corpora and syntactic term patterns. The statistical measure used to extract collocations was the Log-Likelihood Ratio, which was then combined with Inverse Document Frequency (IDF) of single words to determine the "termhood" of candidates. The hybrid version was designed to extract predefined tag patterns regardless of their frequency in the corpus. In both versions, C-value was used to rank nested terms and filter out the terminologically irrelevant ones. Results show that nested collocations are the source of much noise in the statistical version of the system, while in the hybrid version the gain in precision is smaller.

The final results of the term extractor were evaluated by a domain expert and a terminologist, who were also presented with term lists without the removed nested terms. As expected, both system variants scored better after the treatment of nested terms than before. The paper gives an analysis of the results in a comparative Slovene-English setting using two domain-specific corpora.

Regardless of the proven advantages of this approach, some issues concerning term nesting remain subject to discussion as to whether or not they depend on the domain and the target user (see also Estopa 1999). Considering examples such as *steam generator replacement project* we see that the longer collocation is highly significant in a given context, but may be irrelevant for the domain of nuclear engineering in general. The latter, especially for the purposes of traditional terminography, would rather include terms like *steam generator*. We argue that C-value can be used as a valuable tuning parameter in adjusting a term extractor to the user's requirements.

Selected references

Estopa, R. (1999). Extracció de terminologia: elements per a la construcció d'un SEACUSE (Sistema d'Extracció Automatica de Candidats a Unitats de Significació Especializada). PhD Thesis, Universidad Pompeu Fabra, Barcelona.
Frantzi, K.T. in Ananiadou, S. (1996). Extracting nested collocations. In: 16thConference on Computational Linguistics, COLING, p. 41-46.
Nakagawa, H. and Mori, T. (1998). Nested Collocation and Compound Noun for Term Extraction. In: Computerm '98, First Workshop on Computational Terminology, p. 64-71.
Silva, J. F.; Dias, G.; Guilloré, S. ; Lopes, J.G.P (1999): Using LocalMaxs Algorithm for the Extraction of Contiguous and Non-contiguous Multiword Lexical Units. Actes 9th Portuguese Conference in Artificial Intelligence, Springer-Verlag.