Incorporating Context Information for the Extraction of Terms

Katerina T. Frantzi
Dept. of Computing
Manchester Metropolitan University
Manchester, M1 5GD, U.K.
K.Frantzi@doc.mmu.ac.uk

Abstract

The information used for the extraction of terms can be considered as rather 'internal', i.e. coming from the candidate string itself. This paper presents the incorporation of 'external' information derived from the context of the candidate string. It is embedded to the *C-value* approach for automatic term recognition (ATR), in the form of weights constructed from statistical characteristics of the context words of the candidate string.

1 Introduction & Related Work

The applications of term recognition (specialised dictionary construction and maintenance, human and machine translation, text categorization, etc.), and the fact that new terms appear with high speed in some domains (e.g. in computer science), enforce the need for automating the extraction of terms. ATR also gives the potential to work with large amounts of real data, that it would not be able to handle manually. We should note that by ATR we neither mean dictionary string matching, nor term interpretation (which deals with the relations between terms and concepts).

Terms may consist of either one or more words. When the aim is the extraction of single-word terms, domain-dependent linguistic information (i.e. morphology) is used (Ananiadou, 1994). Multi-word ATR usually uses linguistic information in the form of a grammar that mainly allows noun phrases or compounds to be extracted as candidate terms: (Bourigault, 1992) extracts maximal-length noun phrases and their subgroups (depending on their grammatical structure and position) as candidate terms. (Dagan and Church, 1994), accept sequencies of nouns, which give them high precision, but not such a good recall as that of (Justeson and Katz, 1995), which allow some prepositions (i.e. *of*) to be part of the extracted candidate terms. (Frantzi and Ananiadou, 1996), stand between these two approaches, allowing the extracted compounds to contain adjectives but no prepositions. (Daille et al., 1994) also allow adjectives to be part of the two-word English terms they extract.

From the above, only (Bourigault, 1992) does not use any statistical information. (Justeson and Katz, 1995) and (Dagan and Church, 1994) use the frequency of occurrence of the candidate string as a measure of its likelihood to be a term. (Daille et al., 1994) agree that frequency of occurrence "presents the best histogram", but also suggest the likelihood ratio for the extraction of two-word English terms. (Frantzi and Ananiadou, 1996), besides the frequency of occurrence, also consider the frequency of the candidate string as a part of longer candidate terms, as well as the number of these longer candidate terms it is found nested in.

In this paper, we extend *C-value*, the statistical measure proposed by (Frantzi and Ananiadou, 1996), incorporating information gained from the textual context of the candidate term.

2 Context information for terms

The idea of incorporating context information for term extraction came from that "Extended term units are different in type from extended word units in that they cannot be freely modified" (Sager, 1978). Therefore, information from the modifiers of the candidate strings could be used in the procedure of their evaluation as candidate terms. This could be extended beyond adjective/noun modification, to verbs that belong to the candidate string's context. For example, the form *shows* of the verb *to show* in medical domains, is very often followed by a term, e.g. *shows a basal cell carcinoma.* There are cases where the verbs that appear with terms can even be domain independent, like the form *called* of the verb *to call*, or the form *known* of the verb *to know*, which are often involved in definitions in various areas, e.g. ***is known as the singular existential quantifier***, ***is called the Cartesian product.***

Since context carries information about terms it should be involved in the procedure for their extraction. We incorporate context information in the form of weights constructed in a fully automatic way.

2.1 The Linguistic Part

The corpus is tagged, and a linguistic filter will only accept specific part-of-speech sequencies. The choice of the linguistic filter affects the precision and recall of the results: having a 'closed' filter, that is, a strict one regarding the part-of-speech sequencies it accepts, like the N^+ that (Dagan and Church, 1994) use, will improve the precision but have bad effect on the recall. On the other side, an 'open' filter, one that accepts more part-of-speech sequencies, like that of (Justeson and Katz, 1995) that accepts prepositions as well as adjectives and nouns, will have the opposite result.

In our choice of the linguistic filter, we lie somewhere in the middle, accepting strings consisting of adjectives and nouns:

$$(Noun|Adjective)^+Noun \qquad (1)$$

However, we do not claim that this specific filter should be used at all cases, but that its choice depends on the application: the construction of domain-specific dictionaries requires high coverage, and would therefore allow low precision in order to achieve high recall, while when speed is required, high quality would be better appreciated, so that the manual filtering of the extracted list of candidate terms can be as fast as possible. So, in the first case we could choose an 'open' linguistic filter (e.g. one that accepts prepositions), while in the second, a 'closed' one (e.g. one that only accepts nouns).

The type of context involved on the extraction of candidate terms is also an issue. At this stage of this work, the adjectives, nouns and verbs are considered. However, further investigation is needed over the context used (as it is discussed in the future work).

2.2 The Statistical Part

The procedure involves the following steps:

Step 1: The raw corpus is tagged and from the tagged corpus the strings that obey the $(Noun|Adjective)^+Noun$ expression are extracted.

Step 2: For these strings, *C-value* is calculated resulting in a list of candidate terms (ranked by *C-value* as their likelihood of being terms). The length of the string is incorporated in the *C-value* measure resulting to *C-value'*

$$C\text{-}value'(a) = \begin{cases} \log_2 |a| f(a) & |a| = max, \\ \log_2 |a|(f(a) - \frac{\sum_{b \epsilon T_a} f(b)}{P(T_a)}) & otherwise \end{cases} \qquad (2)$$

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

At this point the incorporation of the context information will take place.

Step 3: Since *C-value* is a measure for extracting terms, the top of the previously constructed list presents the higher density on terms among any other part of the list. This top of the list, or else, the 'first' of these ranked candidate terms will give the weights to the context. We take the top ranked candidate strings, and from the initial corpus we extract their context which currently are the adjectives, nouns and verbs that surround the candidate term. For each of these adjectives, nouns and verbs, we consider three parameters:

1. its total frequency in the corpus,
2. its frequency as a context word (of the 'first' candidate terms),
3. the number of these 'first' candidate terms it appears with.

These characteristics are combined in the following way to assign a weight to the context word

$$Weight(w) = 0.5(\frac{t(w)}{n} + \frac{ft(w)}{f(w)}) \qquad (3)$$

where
w is the noun/verb/adjective to be assigned a weight,
n the number of the 'first' candidate terms considered,
$t(w)$ the number of candidate terms the word w appears with,
$ft(w)$ w's total frequency appearing with candidate terms,
$f(w)$ w's total frequency in the corpus.

A variation to improve the results, that involves human interaction, is the following: the candidate terms involved for the extraction of context are firstly manually evaluated, and only the 'real terms' will proceed to the extraction of the context and assignment of weights (as previously).

At this point a list of context words together with their weights has been created.

Step 4: The previously created by *C-value'* list will now be re-ordered considering the weights obtained from step 3. For each of the candidate strings of the list. its context (adjectives, nouns and verbs that surround it) are extracted from the corpus. These context words have either been found at step 3 and therefore assigned a weight, or not. In the latter case, they are now assigned weight equal to 0.

Each of these candidate strings is now ready to be assigned a context weight which would be the sum of the weights of its context words:

$$wei(a) = \sum_{b \epsilon C_a} Weight(b) + 1 \qquad (4)$$

where
a is the examined n-gram,
C_a the context of a,
$Weight(b)$ the calculated (from step 3) weight for the word b.

The candidate terms will be now re-ranked according to:

$$NC\text{-}value(a) = \frac{1}{log(N)} C\text{-}value'(a) \cdot wei(a) \qquad (5)$$

where
a is the examined n-gram,
$C\text{-}value'(a)$ calculated from step 2,
$wei(a)$, the calculated from step 4 sum of the context weights for a,
N the size of the corpus in terms of number of words.

3 Future work

Our future work involves

1. The investigation of the context used for the evaluation of the candidate string, and the amount of information that various context carries. We said that for this prototype we considered the adjectives, nouns and verbs that surround the candidate string. However, could 'something else' also carry useful information? Should adjectives, nouns and verbs all be considered to carry the same amount of information, or should they be assigned different weights?

2. The investigation of the assignment of weights on the parameters used for the measures. Currently, the measures contain the parameters in a 'flat' way. That is, not really considering the 'weight' (the importance) of each of them. So, the measures are at this point a description of which parameters to be used, and not on the degree to which they should be used.

3. The comparison of this method with other ATR approaches. The experimentation on real data will show if this approach actually brings improvement to the results in comparison with previous approaches. Moreover, the application on real data should cover more than one domains.

4 Acknowledgement

I thank my supervisors Dr. S. Ananiadou and Prof. J. Tsujii. Also Dr. T. Sharpe from the Medical School of the University of Manchester for the eye-pathology corpus.

References

Sophia Ananiadou. 1988. A Methodology for Automatic Term Recognition. Ph.D Thesis, University of Manchester Institute of Science and Technology.

Didier Bourigault. 1992. Surface Grammatical Analysis for the Extraction of Terminological Noun Phrases. In *Proceedings of the International Conference on Computational Linguistics, COLING-92*, pages 977–981.

Ido Dagan and Ken Church. 1994. Termight: Identifying and Translating Technical Terminology. In *Proceedings of the European Chapter of the Association for Computational Linguistics, EACL-94*, pages 34–40.

Béatrice Daille, Éric Gaussier and Jean-Marc Langé. 1994. Towards Automatic Extraction of Monolingual and Bilingual Terminology. In *Proceedings of the International Conference on Computational Linguistics, COLING-94*, pages 515–521.

Katerina T. Frantzi and Sophia Ananiadou. 1996. A Hybrid Approach to Term Recognition. In *Proceedings of the International Conference on Natural Language Processing and Industrial Applications, NLP+IA-96*. pages 93–98.

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

Juan C. Sager. 1978. Commentary in *Table Ronde sur les Problémes du Décourage du Terme*. Service des Publications, Direction des Francaise, Montréal, 1979, pages 39–52.