Automatic Term Recognition using Contextual Cues

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

Abstract

In this paper we present an approach for the extraction of multi-word terms from special language corpora. the new element is the incorporation of context information for the evaluation of candidate terms. This information is embedded to the *C-value* method in the form of statistical weights.

1 Introduction

Automatic term recognition (ATR) is the extraction of technical terms from special language corpora with the use of computers. Its applications include specialised dictionary construction and maintenance, human and machine translation, indexing in books and digital libraries, hypertext linking, text categorization etc.

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 conceptsI).

When ATR is concerned with single-word term extraction, domain-dependent linguistic information is used (as in (Ananiadou, 1988), who used morphological information for the recognition of terms in the medical domain of Immunology). Multi-word ATR usually uses linguistic information in the form of a grammar that mainly allows noun phrases to be extracted as candidate terms. The grammar itself may differ: Dagan & Church (1995), accept only sequencies of nouns, which give them a high precision of the output, but not such a good recall as that of Justeson & Katz (1995), which allow some prepositions (ie. *of*) to be part of the extracted candidate terms. Frantzi & Ananiadou (1996a), stand between the above two, allowing adjectives to be part of the noun phrases, but no prepositions. Daille et al. (1994), also allow adjectives to be part of the two-word English terms they treat.

For the statistical element, Justeson & Katz (1995) and Dagan & Church (1995), use pure frequency of occurrence. Daille et al. (1994) agrees that frequency of occurrence "presents the best histogram", but also suggests the likelihood ratio for the extraction of two-word English terms. Frantzi & Ananiadou (1996a), besides total 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 candidate terms.

In this approach a new type of information is incorporated to the approach of (Frantzi & Ananiadou, 1996a): that of the environment[1] of the candidate term. The next section briefly presents terms and particularly multi-word terms, and gives some problems in multi-word ATR. Section 3 briefly talks on *C-value*, and section 4 gives the proposed method, i.e. context information for terms, the linguistic and statistical part, and the algorithm. Conclusions and future work follow.

2 Terms

Terms are the linguistic representation of the concepts in a particular subject field, and "are characterised by special reference" as opposed to words that "function in general reference over a variety of codes" (Sager, 1980). So while the words collectively form the *vocabulary*, the terms of a domain form its *terminology*.

[1] We use the words *context* and *environment* interchangeably.

According to Lauriston (1996), a term is "the intersection between a conceptual realm (a defined semantic content) and a linguistic realm". It is not precise however to talk about 'intersection' unless concepts and termforms are both represented by sets of the same type of elements, and they are not. Then, in order to link term, termform and concept, in a more formal way we say

```
A term T is an ordered pair <c,t>, where c is a concept, from a special
language and t is a termform.
```

From the above we can see that TR is strongly related to TI (term interpretation), and it would be more accurate to talk about *termform recognition*, if we do not involve any TI. However, we use the term *term recognition* in order to be consistent with the previous works (that also do not deal with TI).

Terms are always related to a special language (SL). SL is a language for a restricted type of communication, i.e. medicine, law, mechanical engineering, etc. Briefly we can say here that SLs are based at and derived from the general language (GL), but hold differences in the lexical and semantic level (a GL lexicon for instance would be insufficient if used on a SL text (Ananiadou, 1988; Sager, 1990)).

Most of the difficulties encountered in ATR come from the fact that distinguishing terms from words is not an easy task. Though there exist term formation rules, these are not strong enough to distinguish terms from non-terms.

2.1 Multi-word Terms

Terms may consist of a single wordform so-called *simple* (or *one-word*) terms, or two or more wordforms, called *multi-word* (or *complex* or *extended*) terms.

Multi-word terms have been considered to be the preffered units of designation of terminological concepts. Usually it is the complex relationships that are expressed with multi-word terms, but it is not always the case that the correlation between the complexity of a concept and the length of the term is straightforward (Sager, 1990).

Juxtaposition often indicates terminologisation: *a method for extinguishing fires* is more likely to referring generally, whereas the compound *fire extinguishing methods* is more likely to be a special reference item (Sager, 1978:47). As compounds, the terms result from collocations that are developed into terminological units, with the omission of articles and prepositions: *a stud of clamping* becomes *clamping stud*, *the allocation of space*, *space allocation*, (Sager, 1980:266).

Sager (1980) gives a categorization of the compounds that could serve as multi-word terms in special languages. The most common compounds are the noun compounds, which for terminology purposes includes the adjective-noun and the phrasal compounds. While not very common, compound terms can consist of words belonging to a wide range of parts of speech (Lauriston, 1994)

2.2 Problems in Multi-word ATR

Multi-word ATR presents the following problems:

- Though direct juxtaposition often indicates terminologisation, it does not guarantee it. This is a result of the fact that there are no formal syntactic properties to determine whether a word sequence is a term or not. Moreover, multi-word term structures do not seem to differ from general language structures, and the distinction between them and general language compounds and phrases has not been clearly stated by linguists. English presenting structural ambiguity in parsing makes the problem even bigger: in *Just in time methods have been introduced*, *Just in time* could be either part of the multi-word term *Just in time methods*, or simply a modifier to the verb.

- Another big problem is variation. Variation mainly invludes the following problems:

    1. Hyphenation: the same term can appear in the text with or without a hyphen, or sometimes even as a simple (one-word) term: *tool box, tool-box, toolbox.* Even dictionaries are not systematic in the use of hyphens (Sager, 1980:266).
    2. Abbreviation: for economy, the term is used with the omission of one or more words: *gearbox end cover plate, end cover, cover.*

3 C-value

The *C-value* statistical measure for the extraction of multi-word terms was described in (Frantzi & Ananiadou, 1996a). Here we briefly review its features.

The procedure for extraction starts with the strings of maximum length. So, if for instance we decide that the longer strings we want to extract are of length n, it will start with them. In that case the only parameter involved in their likelihood for being candidate terms, is their frequency in the corpus. So, if a is the candidate string, and $f(a)$ its frequency,

$$C\text{-}value(a) = f(a) \tag{1}$$

Then comes the extraction of the directly shorter strings. For each of them (and for the next steps which will every time extract the directly shorter strings), three parameters are considered:

1. the string's total frequency of occurrence in the corpus,
2. its frequency of occurrence in longer (already extracted) candidate terms,
3. the number of these longer candidate terms.

The first parameter is due to the fact that technical terms tend to appear with high frequencies (however, a high frequency is not a guarantee of termhood, and vice versa, a term does not always appear with a high frequency). The second and third factors are in order to prevent substrings of terms to be erroneously extracted as terms due to their 'high' frequency of occurrence. As an example consider the following:
soft contact lenses
hard contact lenses
contact lenses
soft contact
If we only use the frequency of the candidate string, then if *soft contact lenses* was extracted as a candidate term, *contact lenses* and *soft contact* would be also extracted since they present at least the frequency of *soft contact lenses* (*contact* is tagged as a noun, so it passes through the filter). Now, while *contact lenses* should be extracted as a candidate term, *soft contact* should not. This is where the second and third factors are involved.

The claim is that a substring of a candidate term is a candidate term itself, if it shows 'sufficient' independence from the longer candidate terms it appears as a substring of. This independence is measured as a function of the frequency by which it appears in longer candidate terms, and to the number of these longer candidate terms. So, while a high frequency of a candidate string in longer candidate terms is a minus, if the number of these longer terms is big, the substring shows independence, which is a plus.
The above are combined into the following measure

$$C\text{-}value(a) = f(a) - \frac{t(a)}{c(a)} \tag{2}$$

where
a is the examined string,
$f(a)$ is the total frequency of occurrence of a on the corpus,
$t(a)$ is the frequency of occurrence of a in longer (already extracted) candidate terms.
$c(a)$ is the number of those candidate terms.

(We must keep in mind that the above only describes likelihood. The measure extracts a list of candidate terms whose final evaluation is to be done manually).

4 Context information for terms

The environment of words has been previously used for the construction of thesaurus (Grefenstette, 1994). In that case, words that share the same context are viewed as synonymous.

Regarding terms, the idea of incorporating context information for the extraction came from the fact that extended word units can be freely modified while multi-word terms cannot (Sager, 1978). So,

information that could be used in the procedure for the assignment of a value to candidate terms, could be gained from their modifiers. This could be extended beyond adjective or noun modification, to verbs that belong to the candidate term's context. For example, the form "shows" of the verb "to show" in medical domains, is almost always followed by a term, e.g. *shows a basal cell carcinoma*. There are cases where the context that appears 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 the extraction of terms.

We incorporate this type of information to the approach of Frantzi & Ananiadou (1996a) for the extraction of multi-word terms.

The way we choose to assign the weights on the context is fully automatic, and the procedure can be briefly described as follows:

1. Produce a list of candidate terms using the *C-value* approach.
2. Get some of the 'first' strings of the produced list. These 'first' strings present the higher density in terms of the whole produced list.
3. Extract the context for the above 'first' candidate terms from the corpus. We consider this context to be the verbs, adjectives and nouns that surround the candidate term.
4. According to some of their statistical characteristics, that we will discuss later, assign to each of those verbs, adjectives and nouns, a weight.

4.1 The Linguistic Filter

The corpus used is tagged, and a linguistic filter will only permit specific part-of-speech strings to be considered. The choice of the linguistic filter affects the precision and recall of the results. So, having a 'closed' filter, that is, one that does not allow 'many' part-of-speech sequencies, like the N^+ that Dagan & Church (1994) use, will improve the precision but have negative effect on the recall. On the other side, an 'open' filter, one that allows more part-of-speech sequencies, like that of (Justeson and Katz; 1995), that allows prepositions as well as adjectives and nouns, will have the opposite result.

Our choice of the linguistic filter lies somewhere in the middle, allowing strings consisting of adjectives and nouns:

$$(Noun|Adjective)^+Noun \tag{3}$$

However, we do not claim that this specific filter should be used at all cases, but that its choice could be either more 'closed' or 'open' depending on the application: the construction of domain-specific dictionaries would 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 can be quick. 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 and adjectives).

The type of context appropriate that characterises a term is also involved in the linguistic element. At this stage of our work, we consider the verbs, adjectives and nouns. However, further investigation will take place to refine the context used.

4.2 The Algorithm

The following stages take place:

1. The raw corpus is tagged with Brill's part-of-speech tagger (Brill, 1992). From the tagged corpus the n-grams that obey the $(Noun|Adjective)^+Noun$ expression are extracted.
2. For these n-grams, *C-value* is calculated resulting on a list of potential terms ranked by *C-value* (as their likelihood of being terms). In this use of *C-value*, the parameter of the length of the n-gram is incorporated. The length had been previously considered when *C-value* was used for

the extraction of collocations (Frantzi & Ananiadou, 1996b), but not for the extraction of terms. We weaken the length weight, and obtain *C-value'*:

$$C\text{-}value'(a) = \begin{cases} \log_2 |a| \cdot f(a) & |a| = max, \\ \log_2 |a| \cdot (f(a) - \frac{1}{c(a)} \sum_{i=1}^{c(a)} f(b_i)) & otherwise \end{cases} \quad (4)$$

where
a is the examined n-gram,
$|a|$ the length, in terms of number of words, of a,
$f(a)$ the frequency of a in the corpus,
b_i the candidate extracted terms that contain a,
$c(a)$ the number of those candidate terms.

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

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. So we take the top ranked candidate strings, and from the initial corpus extract their context (or else their corcondances) which currently are the verbs, adjectives and nouns that surround the potential term. For each of these verbs, adjectives and nouns, we consider three parameters:

   (a) its total frequency in the corpus,

   (b) its frequency as context word (of those 'first' n-grams),

   (c) the number of those n-grams it appears with.

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

$$Weight(w) = 0.5 \cdot (\frac{t(w)}{n} + \frac{ft(w)}{f(w)}) \quad (5)$$

   where
   w is the noun/verb/adjective to be assigned a weight,
   n the total number of 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 that are involved for the extraction of context are first evaluated, and only the 'real terms' will proccede for the extraction of the context, and the assignment of weight to it.

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

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

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

$$wei(a) = \sum_{b \epsilon C_a} Weight(b) + 1 \quad (6)$$

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

The n-grams will be now reranked according to:

$$NC\text{-}value(a) = \frac{1}{log(N)} \cdot C\text{-}value'(a) \cdot wei(a) \quad (7)$$

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

Table 1 shows the first 80 candidate terms of the produced list.

5 Conclusions & Future work

This paper gives an approach of incorporating context information for the extraction of multi-word terms. Till now context information has been used for the extraction of synonymous words, while for the extraction of terms, the information to be used was rather 'internal', that is, linguistic and statistical that characterised the candidate term and not its environment. For the current implementation, there has not been complete investigation on the type of context to be considered. The verbs, adjectives and nouns that surround the candidate term are used, all assumed to carry the same amount of information.

The future work involves the following:

- The investigation of the context used for the evaluation of the candidate strings, and the amount of information that various context carries. In this paper we have considered the verbs, adjectives and nouns to give us information about the candidate term, but could it be something else as well? Do verbs, adjectives and nouns, all carry the same amount of information, or should they be assigned weights according to their part-of-speech?
- The investigation of the assignment of weights on the parameters used for the measures. Till now the parameters were used in a rather *flat* way.
- The comparison of this method with other ATR approaches, by applying them on the same data that should cover more than one domains.

6 Acknowledgements

We thank Dr. Tom Sharpe from the Medical School of the University of Manchester, for providing us with the corpus.

References

Ananiadou, S., 1988. A Methodology for Automatic Term Recognition. PhD Thesis, University of Manchester Institute of Science and Technology.

Bourigault, D., 1992. Surface Grammatical Analysis for the Extraction of Terminological Noun Phrases. In *Proceedings of COLING*, 977–981.

Brill, E., 1992. A simple rule-based part of speech tagger. In *Proceedings of the 3rd Conference on Applied Natural Language Processing, ACL*, 152–155.

Dagan, I. and Church, K., 1994. Termight: Identifying and Translating Technical Terminology. In *Proceedings of EACL*, 34–40.

Daille, B.; Gaussier, E. and Lange, J.M., 1994. Towards Automatic Extraction of Monolingual and Bilingual Terminology. In *Proceedings of COLING 94*, 515–521.

Frantzi K. and Ananiadou S., 1996a. A Hybrid Approach to Term Recognition. In *Proceedings of NLP+IA*, 93–98.

Frantzi K. and Ananiadou S., 1996b. Extracting Nested Collocations. In *Proceedings of COLING*, 41–46.

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

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

Lauriston, A. 1994. Automatic recognition of complex terms: Problems and the TERMINO solution. In *Terminology*, 1:147-170.

Lauriston, A. 1996. Automatic Term Recognition: Performance of Linguistic and Statistical Learning Techniques. PhD Thesis, University of Manchester, Institute of Science and Technology.

Sager, J.C., 1990. A Practical Course in Terminology Processing. John Benjamins Publishing Company.

Sager, J.C.; Dungworth, D.; McDonald P. F., 1980. English Special Languages. Oscar Brandstetter Verlag KG - Wiesbaden.

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

C-value	String	*C-value*	String
28714.7	BASAL CELL	177.789	IRIS STROMA
24038.6	OPTIC NERVE	174.174	SPINDLE CELLS
10262.8	FIBROUS TISSUE	172.261	CORNEAL STROMA
7352.08	BASAL CELL CARCINOMA	171.224	NAEVOID CELLS
4081.44	ANTERIOR CHAMBER	169.143	NASAL SIDE
3955.45	CELL CARCINOMA	166.31	PARS PLANA
2707.57	TRABECULAR MESHWORK	158.381	STRIATED MUSCLE
2599.49	OPTIC NERVE CUT	150.415	AXIAL REGION
1494.28	SUBSTANTIA PROPRIA	142.195	KERATOTIC DEBRIS
1307.48	GIANT CELLS	140.142	WHITE TISSUE
1153.3	CORNEAL DIAMETERS	135.789	GREY NODULE
1037.78	CILIARY PROCESSES	130.193	COLLAGENOUS TISSUE
972.478	LENS CAPSULE	127.854	WHITE EXCRESCENCE
874.222	HYALINE FIBROUS TISSUE	119.77	SCAR TISSUE
791.822	GREY TISSUE	114.881	OCULAR STRUCTURES
744.72	RETINAL SPACE	114.06	GREYISH WHITE
698.034	TUMOUR CELLS	111.176	SCAR TRACK
623.793	BASALOID CELLS	104.424	LASH LINE
547.523	RETINAL DETACHMENT	101.749	GREYISH TISSUE
522.461	KERATINOUS DEBRIS	99.1182	WHITE NODULE
497.151	PLASMA CELLS	98.0175	SQUAMOUS EPITHELIUM
462.712	KERATINOUS CYST	94.2403	NODULAR EXCRESCENCE
461.63	PUPILLARY BORDER	89.6127	SWEAT DUCT
456.187	NERVE HEAD	89.4159	OCULAR HAEMORRHAGE
451.231	LID MARGIN	88.3705	SURGICAL EXCISION
451.125	NAEVUS CELLS	87.7137	BASALOID PAPILLAE
444.721	BULLOUS SEPARATION	85.7369	MITOTIC FIGURES
405.602	CELL PAPILLOMA	82.39	LENS FIBRES
325.321	BASAL CELL PAPILLOMA	79.3124	RETINAL FUNNEL
307.776	CORNEAL EPITHELIUM	79.1997	SEROUS DETACHMENT
303.166	SPINDLY CELLS	77.8568	RETINAL COAGULUM
301.651	FIBROUS STROMA	74.3617	SCLERAL LACERATION
273.895	LYMPHOCYTIC INFILTRATION	73.9586	FATTY TISSUE
272.396	CONJUNCTIVAL EPITHELIUM	72.6021	VASCULAR TISSUE
256.449	TEMPORAL SIDE	69.2714	CONNECTIVE TISSUE
217.739	GIANT CELL	64.9845	DIAMETER X
214.514	BLOOD VESSELS	63.9308	CORNEAL DISC
212.902	BASAL CELLS	59.28	OPTIC NERVE HEAD
205.472	SCLERAL EXTENSION	57.5563	OVERLYING EPIDERMIS
203.519	RED CELLS	57.5465	LENS REMNANT
184.022	CELLULAR FIBROUS TISSUE	56.1416	GOBLET CELLS

Table 1: The first 80 n-grams extracted.