Evaluation of Automatic Term Recognition of Nuclear Receptors from MEDLINE

Sophia Ananiadou[1] s.ananiadou@salford.ac.uk
Sylvie Albert[2] albert@lionbioscience.com
Dietrich Schuhmann[2] Dietrich.Shuhmann@lionbioscience.com

[1] Computer Science, School of Science, University of Salford, The Crescent, Salford, Greater Manchester, M5 4WT, United Kingdom
[2] LION bioscience AG - Post Box 10 37 80 - D.69027 Heidelberg, Germany

Keywords: automatic term recognition, C/NC-value approach

1 Introduction

In this paper we examine the use of the C/NC value approach for the automatic extraction of terminological units from MEDLINE corpora as used in the BioPath project. We briefly present our approach, the methodology adopted and preliminary results.

2 Method and Results

The C/NC value method combines linguistic and statistical information for the automatic extraction of multi-word technical terms from machine readable corpora. C-value enhances the commonly used statistical measure of frequency of occurrence for term extraction, making it sensitive to a particular type of multi-word terms, nested terms. Nested terms are those which also exist as substrings of other terms. Consider the following term: *mitogen-activated protein kinase*. Valid substrings of the longer term are also extracted, e.g. *mitogen-activated protein*, *protein kinase*. NC/value incorporates contextual information in the form of statistical (weights) and linguistic information improving the performance of C/value. Deeper forms of contextual information (semantic knowledge) have also been used. The measures have been applied to medical corpora in English and have been also adapted for Japanese.

We applied the C/NC value into a collection of 2000 abstracts from MEDLINE. The domain area was that of nuclear receptors.

Before applying the first measure (C-value) we tagged the corpus using MXPOST tagger, freely available from the University of Pennsylvania. The tagger was not trained for our corpus. The wordforms were stemmed using Porter's stemming algorithm. Stemming was used to deliver better results for statistical measures. The same corpus was run on a different tagger based on constraint based grammars, ENGCG.

We used a domain specific stop list which was produced after the first results of C/NC value, containing single words (e.g. dramatic, data) and multiword units (e.g. brief review).

The following linguistic filters were applied, based on commonly used term formation patterns:

1. Noun + Noun
2. (Adj | Noun) + Noun
3. (Adj | Noun) + | ((Adj | Noun) * (NounPrep) ?) (Adj |Noun)*) Noun

We performed experiments using the 3 filters, with and without stemming, using both taggers. Initial results show a better performance using a closed filter (1 & 2), using stemming, based on the tagged results of MXPOST. Due to space limitations we do not provide the evaluation using ENGCG with different filters and stemming.

The results of NC value, using MXPOST, filters 1 & 2 and stemming, are given in Table 1.

Table 1:

NC value	Term form
127.44	Protein kinas
85.31	Growth factor
57.41	Transcript factor
42.96	Gene express
42.74	Protein kinas C
39.83	Annexin II

Stemmed forms were mapped to possible surface forms which were then associated with the stemmed forms. The mapping was done in a separately programmed process after the Ncvalue programs had run.

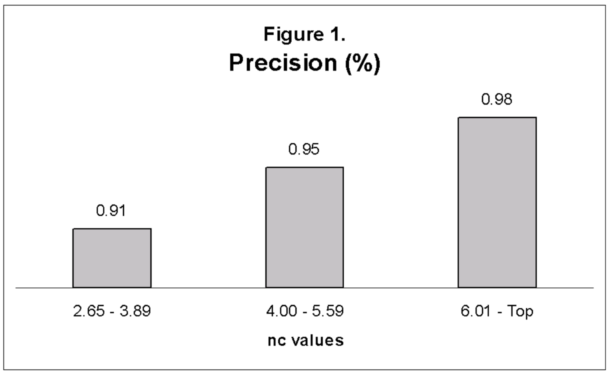
In table 2 wee see the stemmed form (NC value) and the set of matched surface forms.

Table 2:

STEM:	amino acid sequenc (17.95)
	Amino acid sequence
	Amino acid sequences
	Amino acid sequencing

3 Evaluation

The results of NC/value were mapped into possible surface forms. Domain experts evaluated the results of the mapped forms. Evaluation was performed in terms of precision. The distribution of terms in the extracted list (NC/value) demonstrated a very high precision for the top ranked terms, the precision decreasing slightly as we reach the bottom level. We evaluated precision using 3 intervals. Recall is the same for NC/value as for C/value. The Precision, evaluated for the top three ranges of NC/value, is shown in Figure 1.



References

[1] Frantzi, K. and Ananiadou, S., The C-value/NC-value domain independent method for multiword term extraction, *Journal of Natural Language Processing*, 6(3):145–179, 1999.

[2] Maynard, D. and Ananiadou, S., Identifying Terms by their Family and Friends, *Proceedings of the 18th International Conference on Computational Linguistics*, 530–536, 2000.

[3] http://open.muscat.com/stemming/

[4] http://www.cis.upenn.edu/ adwait/statnlp.html

[5] ENGCG English Constraint Grammar http://www.lingsoft.fi/