An ATE system based on probabilistic relations between terms and syntactic functions

Xing Zhang, Alex Chengyu Fang

Department of Chinese, Translation and Linguistics, City University of Hong Kong,
Tat Chee Avenue, Kowloon, Hong Kong

Abstract

This research aims at constructing an Automatic Term Extraction (ATE) system by exploiting extended syntactic information. More specifically, the ATE system will be built on the relations between syntactic functions and term occurrence. Thus, after training on a large medical corpus drawn from MEDLINE, the proposed ATE system will find out which syntactic functions are good indicators for terms. Accordingly, term candidates occurring in these syntactic structures will be assigned a higher weight for better probabilistic estimates. In this way, syntactic function information will be used to measure the termhood of term candidates. The proposed ATE system focuses on syntactic features that will be helpful in identifying terms. It proposes one linguistic method, SF-Value, to weight the termhood of term candidates. The most innovative aspect of this research is to explore the role of syntactic functions in recognizing and extracting terms from texts. This approach represents a novel, linguistically motivated perspective in the area of terminological processing. Most importantly, unlike ATE systems that include various terminology extraction techniques, this system lies mainly on syntactic properties of terms. This emphasis may not as effective as other techniques in terms of utilizing features of term candidates, but it makes an effort to explore how syntactic information of term candidates can be helpful in this regard. Furthermore, this system evaluates its results by comparing them with those obtained with the C-Value method. The overall performances of these two methods are slightly different. Though little improvement is made, this system also attempts to combine C-Value and SF-Value together to refine term extracting results. The most significant aspect of this research lies in that it explores the contribution of syntactic information to term extraction.

Keywords: term extraction, syntactic function, syntactic structure, termhood

1. Introduction

Terms usually refer to the linguistic manifestation of concepts in a specific domain. More specifically, terms are the linguistic expression of the concepts of special communication and are organized into systems of terms which ideally reflect the associated conceptual system (Ananiadou, 1994). In the current paper, a term is defined as a word or phrase that denotes specific concepts in a given domain and is categorized as a term by terminologists or experts.

ATE has recently gained great interest of researchers in Natural Language Processing as it is much needed to build a coherent terminology and serves as the starting point of many applications such as human or machine translation, indexing, thesaurus construction, knowledge organization, document retrieval, and summarization. Many efforts have been made to improve the automatic term extraction process by means of different kinds of techniques. However, the past approaches to ATE have been largely based on statistic information with limited efforts to explore the role of linguistic information in selecting and ranking term candidates. Considering

syntactic functions as an important factor, this study attempts to explore the contribution that syntactic information will add to ATE.

Most importantly, unlike ATE systems that include the best reported terminology extraction techniques (*e.g.* Sclano and Velardi, 2007), the proposed system lies mainly on syntactic properties of terms. This emphasis may not be the most effective techniques in terms of selecting term candidates, but it makes an effort to explore how syntactic information of term candidates can be helpful in this regard.

1.1. Techniques in ATE system

In general, the techniques for ATE can be categorized into three approaches as linguistic, statistical and hybrid. Linguistic approach is based on linguistic pre-processing and annotations, such as taggers and shallow parsers. Moreover, stop-lists and term variations can be taken into consideration as further refinement. Linguistic approaches usually take advantage of linguistic cues such as morphological analysis, part of speech, grammatical structure of possible terms, lemmatization for inflected forms and tokenization to identify word and sentence boundaries. This approach is now often used as the preparation for a statistical approach, that is, it identifies possible terms, and then a statistical approach is applied in order to calculate the termhood of the term candidates. After linguistic filtering, various measures are employed in the literature for grading the termhood/collocativity of collected candidates.

Some early ATR systems are representative in terms of their handling of linguistic features in their approaches, such as LINGSOFT, FASTR (Jacquemin, 1996) and LEXTER (Bourigault, 1992). Some classic statistical measures used in ATE are TF·IDF (Salton and Buckley, 1988), Mutual Information (Church and Hanks, 1989), Log-Likelihood Ratio (Dunning,1993), Dice Factor (Smadja et al., 1996), etc. The most commonly used one is C-value, which can be said as a statistical measure. It assigns termhood value to a candidate string, ranking it in the output list of candidate terms. As pointed out by many studies such as (Wermter et Hahn, 2005; VU et al., 2008) that the measure C-value is method (Frantzi et al., 1998) that widely considered as the state-of-the-art model for ATE. Although this method was first applied on English, it also performed well on other languages such as Japanese (Mima and Ananiadou, 2001), Slovene (Špela Vintar, 2004), and other domains such as medical corpus (Frantzi et al., 1998), and computer science (Milios et al., 2003). As pointed out by many studies such as (Wermter and Hahn, 2005; VU et al., 2008) that the measure C-value is method (Frantzi et al., 1998) is widely considered as the state-of-the-art model for ATE.

The formula for C-value is as

$$\text{C-value (a)} = \begin{cases} \log 2|\alpha| \bullet f(a) & \text{a is not nested,} \\ \log 2|\alpha| \left(f(a) - \frac{1}{P(Ta)} \sum b \in T_a f(b)\right) & \text{otherwise.} \end{cases}$$

Where a is the candidate string, f(a) is its frequency of occurrence in the corpus, T_a is the set of extracted candidate terms that contain a, $P(T_a)$ is the number of these candidate terms.

Hybrid methods occupy a major position in ATE and it's the general trend. Traditionally, hybrid approaches make use of a statistical system that includes some linguistic information. The most representative method is NC-Value (Frantzi et al. 1998), which combines linguistic and statistical information with an emphasis on nested terms by enhancing common statistical measure of frequency of occurrence.

However, information exploited to help select terms is often centered on statistic properties of terms or some basic linguistic features, or corpus comparison. There lacks deeper syntactic research to recognize terms. Recently, a few of researchers have moved from exclusive preoccupation with noun phrases to surrounding verbs, which arises from the verb's traditional role as the central organizer and distributor of concepts in a sentence. Some studies have analyzed the role of verbs in indicating occurrence of term, either in the microstructure of discourse, such as verb-argument relations (Amig'o et al., 2004; Eumeridou et al., 2002; Eumeridou et al., 2004), or in macro discoursal structure (Teruo and Kyo, 2004).

1.2. MEDLINE

Typically most of these studies on term extraction evaluate the goodness of their algorithms by consulting domain experts who identify whether a ranked candidate is a true term or not. However, even for domain experts, it may not be so easy to decide what a relevant term is in a particular domain. If you consult different experts, the results will possibly be different. Therefore, instead of relying on solely human judgment, a more objective solution may make use of already existing terminological resources.

Based on this consideration, more and more ATE systems have chosen MEDLINE as the test bench (Wermter and Hahn, 2005). Wermter and Hahn (2005) build the measure of termhood on the limited paradigmatic modifiability of terms and test the corresponding algorithm on bigram, trigram and quadgram noun phrases extracted from MEDLINE.

MEDLINE is the National Library of Medicine's premier bibliographic database. It is the world's most comprehensive source of life sciences and biomedical bibliographic information. MeSH is the U.S. National Library of Medicine's controlled vocabulary used for indexing articles for MEDLINE. MeSH terminology provides a consistent way to retrieve information that may use different terminology for the same concepts. There are 24,767 descriptors in 2008 MeSH. There are also over 97.000 entry terms that assist in finding the most appropriate MeSH Heading (Retrieved 15 July 2008 from MeSH website).

2. Methodology

This study proposes a method to measure the probabilistic relations between terms and their syntactic functions. The results are used to calculate termhood of candidates according to their syntactic functions. Overall, this system includes three major models. The first model is to get abstracts from MEDLINE. It will create experimental corpus from MEDLINE database. The second model creates a term list from MeSH, and annotates the corpus according to this term list. Another major function of the second model is to compute the term occurrences in different syntactic structures. The third model uses the knowledge of term rates in different structures to assign different weights to different syntactic structures and then compute the syntactic function value of each NP using the following formula:

$$SFValue = \sum_{i=1}^{n} FSS_i \times WSS_i$$

SF-Value is the syntactic function value of an NP, FSS is the frequency of syntactic structure$_i$, WSS_i is the weight of syntactic structure$_i$, n is the count of how many syntactic structures this term candidate occurs in.

WSS_i is computed proportionally according to term rates in this syntactic structure after training on three sub-corpora. For example, terms occurred in NPs of the function Subject account for

24.28% among all the terms in training sub-corpus 1, and then the WSS of this function in this corpus is 0.2428. If the WSS of this function in training sub-corpus 2 and sub-corpus 3 are 24.53% and 24.03% respectively, the final WSS of the function Subject would be based on the average value of these three rates (i.e. 24.28%, 24.53% and 24.03%).

3. Resource Building and Processing

3.1. Corpora Building up

This paper proposes to build a small subset of MEDLINE abstracts based on the controlled search of the database using the keyword *internal medicine*. This search produces nearly 252.033 abstracts (until 17 July 2008). Each abstract consists of a single title and a number of sentences. Four sub-corpora were created, totaling 339.555 words in 19.949 sentences. The first three of them are training corpora; the fourth one will be used as testing corpus. All of them are first parsed by the Survey Parser (Fang, 1996), which can produce syntactic functions for every unit. And the first corpus is also manually checked by human professionals.

The first sub-corpus contains 5.000 parsed sentences, the second one contains 4.821 parsed sentences, and the third one contains 4.125 parsed sentences. The fourth corpus for testing contains 4.114 parsed sentences. A list of medical terms was created from Medical Subject Headings (MeSH) beforehand. These corpora then will be terminologically annotated: noun phrases that match the term list are tagged as terms. The following table shows the basic information of the related data in the experiment.

	Training corpus			*Testing corpus*
	1st sub-corpus	*2nd sub-corpus*	*3rd sub-corpus*	*4th sub-corpus*
Abstracts	365	404	402	334
Sentences	5,000	4,821	4,125	4,114

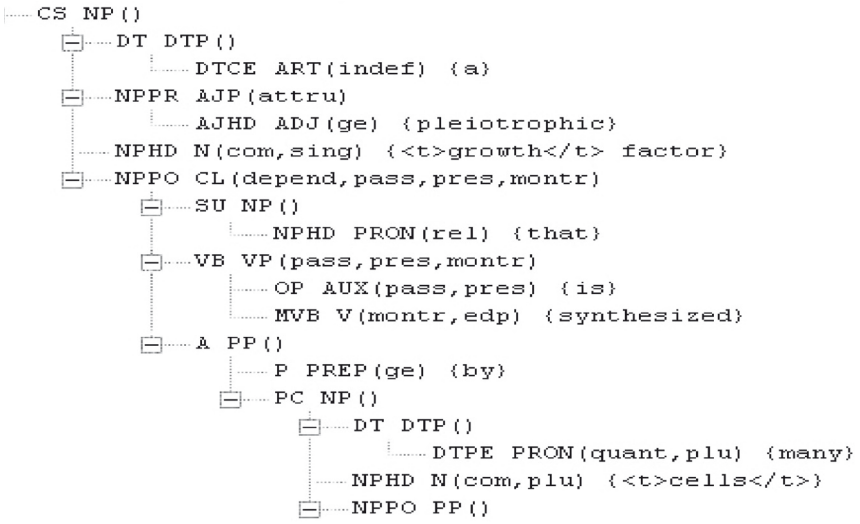
Table 1: Basic data of resource

3.2. Survey Parser

Survey Parser was first designed to complete the syntactic annotation of the International Corpus of English. It effectively parses sentences in many layers with detailed syntactic functions. The unique complexity of the parsing scheme is that it analyzes the syntactic functions of the constituent structures and represents them in the form of a parsing tree. The following examples show some of the phrases and internal structures in the parsing tree after annotated with terms from MeSH term list (see Pic. 1):

Survey Parser also classifies syntactic functions into two major kinds: one is phrasal functions, and the other is clausal functions, which correspond to the basic elements of English sentences such as subject, verb, object, complement and adverbial. In current study, an effort is made to make use of the syntactic functions in three ways which include immediate clausal functions, immediate mother nodes and conflated clausal functions. So, in a syntactic structure used to describe where a term occurs, the ending node may be an immediate clausal function or an immediate mother node or the conflation of the same ending clausal function. Clausal syntactic functions in Survey Parser (Fang, 2000): include subject (SU), provisional (PRSU), notional (NOSU), subject (SU), provisional (PRSU), notional (NOSU), direct (OD), indirect

(OI), provisional (PROD), notional (NOOD), (CS), object (CO), transitive (CT), focus (CF), adverbial (A) etc.



Picture 1: Parsed and term annotated trees from survey parser

Take the above tree as example, *cells* is tagged as a term, then, the syntactic structure for it is "NPHD-N%PC-NP%A-PP". This study adopts all the symbols used in the Survey Parser, for example, SU stands for subject, PC stands for prepositional complement and so on. And '%' is used to indicate a node of higher level. '+' is used to indicate two nodes of the same level.

In this experiment, a stop word list is created to consist of a few frequent grammatical words, such as definite articles, demonstrative and possessive adjectives, and indefinite articles. Words in stop list are uninformative for terminology extraction. The aim of using a 'stop word' list is to remove very frequent words which are not considered to carry terminological meanings.

4. System Design

Combing the above factors together, the proposed extractor will be implemented in two modules. Moreover, as the C-Value is considered as the most effective statistical means in ATE, the current system will implement C-Value as well to compare the result with that from the SF-Value proposed in this paper. The first module is to compute term rates in different syntactic functions, which includes the following steps:

- Have training and testing texts parsed by the Survey parser.
- Input training texts and extract all the NPs with their frequencies from it.
- Use stop list to filter those NPs that are impossible to be terms; and delete those with a length larger than10 words. The system produces a NP list.
- Compute C-Value for all the NPs in the list, and arrange those NPs in descending order.
- Compute the terms rates in different syntactic structures, and compute different weights (WSS) respectively for those syntactic structures.

The second module is for processing testing corpus and evaluation of results.

- Input testing texts and extract all NPs with their frequencies from it.
- Produce a NP list and filter this list with stop list.
- Compute the frequencies of these syntactic structures for each NP where this NP has occurred.
- Compute SF Value for each NP, and arrange them in descending order.

- Based on the minimum value and maximum value of C-Value and SF-Value, automatically set optimum thresholds for these two values to produce the highest F-score value.

$$F - score = 2\frac{(precision \times recall)}{precision + recall}$$

- Compare the F-score value of C-Value and SF-Value.
- Combine the methods of C-Value and SF-Value and evaluate its performance.

5. Results of Term Rates in Syntactic Structures

In the first module, we can find a roughly consistent tendency in respect of the ordering of term rates of syntactic structures across the three sub-corpora (Tab. 2). Especially, if we group these syntactic structures with the term rates below 25% and above 2% as the first group, we can see the ranking of these syntactic structures are exactly the same. With regard to the second group, their values fall between 2% and 1%, but these structures are not ranked in the same order. In this system, they are considered as having the same SWW as there is not much difference between their values.

Group 1	*Syntactic Functions*	*Sub-corpus 1*	*Sub-corpus 2*	*Sub-corpus 3*
1	SU-NP	24.28%	24.53%	24.03%
2	PC-NP%A-PP	19.38%	19.16%	19.22%
3	OD-NP	10.20%	8.81%	9.30%
4	PC-NP%NPPO-PP%PC-NP%A-PP	6.04%	6.61%	5.93%
5	PC-NP%NPPO-PP%SU-NP	3.68%	3.57%	3.51%
6	PC-NP%NPPO-PP%OD-NP	2.18%	2.00%	2.07%
Group 2	*Syntactic Functions*	*Sub-corpus 1*	*Sub-corpus 2*	*Sub-corpus 3*
7	A-PP	1.97%	1.84%	1.84%
8	NPPR-AJP%SU-NP	1.44%	1.28%	1.24%
9	NPPR-AJP+NPHD-N%SU-NP	1.23%	1.46%	1.27%
10	NPPR-AJP+NPHD-N%PC-NP%A-PP	1.19%	1.24%	1.17%
11	CS-NP	1.10%	1.37%	1.49%

Table 2: Term occurrences rates in immediate clausal functions

In Tab. 2, we can find more than three hundred of syntactic structures if we choose clausal functions as ending nodes. However, there are only a few structures accounting most prominently, such as SU-NP, PC-NP%A-PP, while the rest has quite a low frequency each. For example, we can see terms take the function of *subject* most frequently, while these taking the function of *object* account nearly half of the *subject*. The third one is the function *adverbial*. And others may form different phrases with different functions before serving as clausal functions as a whole. Therefore, this ATE system has to deal with such sparse data while trying to keep these possible distinctive features of these syntactic functions.

Besides these major syntactic structures in these two groups, there are several hundreds of structures that terms occur less than 5 times each. Some of them even account for as low as 0.005%. Therefore, these structures with such lower frequencies are conflated before allocating weights to them. The principle for conflation is syntactic structures with the same beginning syntactic functions and the same ending clausal functions are conflated. For example:

Syntactic structures	*Term Rates*
NPPR-AJP+NPHD-N%PC-NP%NPPO-PP%*SU-NP*	0.199%
NPPR-AJP+NPHD-N%PC-NP%NPPO-PP%PC-NP%NPPO-PP%*SU-NP*	0.003%
NPPR-AJP+NPHD-N%NPPO-PP%NPPO-PP%NPPO-PP%*SU-NP*	0.001%

Table 3: Conflation of syntactic structures

In Tab. 3, the starting syntactic functions of these three syntactic structures are all *NPPR-AJP+NPHD-N*, which means a node of NPPR-AJP together with a node of NPHD-N. And the ending syntactic functions are all *SU-NP*, therefore, these three structures are conflated as *NPPR-AJP+NPHD-N*%*SU-NP*, and the term rates of them is added up as 2.03% after conflation.

6. Evaluation

In order to have a basic idea of difference between these two methods, an evaluation to compare these two methods, C-Value and SF-Value are conducted first. Most importantly, the objective of this study is to find a way to combine C-Value and SF-Value together to improve ATE performance. First, it sets a threshold C-Value to rule out NPs whose C-Value is below this threshold and produces a basic term candidates list. And then it computes SF-Value for all term candidates in this list. Moreover, it sets a threshold SF-Value and removes from this list those term candidates whose SF-Values are below it. The system will exhaustively test all the possible threshold C-Values and threshold SF-Values to get the optimal F-score.

Tab. 4 shows the results obtained from these two methods. For these NPs extracted, the highest C-Value is 7.841, and the lowest is 0.334. When the threshold of C-Value is set as 0.9, the maximum F-Score will be 0.420, as the recall is 0.697, precision is 0.299. After testing all the possible threshold values of C-Value, the highest recall we can get is 0.894, and the highest precision is 0.445.

As for the SF-Value of these NPs, the highest is 36.443, the lowest is 0. When the threshold of SF-Value is set as 0.1, we can get the maximum F-score, which is 0.410, as the recall is 0.600, precision is 0.308. Again, after testing all the possible threshold values of SF-Value, the highest recall we can get is 1, and the highest precision is 0.435.

	precision	*recall*	*F-score*
C-Value	0.445	0.894	0.420
SF-Value	0.435	1	0.410
C-Value & SF-Value	0.458	0.894	0.420

Table 4: Comparison between C-Value, SF-Value and C-Value & SF-Value Combined

From the results in the above chart, we can see the recall of the SF-Value is 0.106 higher than that of C-Value while the precision of the C-Value is 0.01 higher than that of SF-Value. And the difference between the F-scores of these two values is 0.01%, which is extremely slight. Therefore, these two methods can be said that they do not have too much different performance at this stage.

While after combining C-Value and SF-Value, the maximum precision comes to 0.458, higher than either C-Value or SF-Value used alone. And the highest recall is 0.894, the same as that of C-Value. Moreover, the optimal F-score is 0.420 (as recall=0.695, precision=0.300), which is also the same as that of C-Value.

7. Conclusion

This study sets out to examine whether syntactic knowledge can be effective in weighting the term status in a specific domain. Based on parsed trees by the Survey parser, a robust parsing system that applies a syntactically rich annotation scheme to natural texts, the proposed system examines the term rates in different syntactic structures. Then it assigns weights to term candidates accordingly if they occur in relevant syntactic structure. Therefore, this feature promotes us to obtain reliable statistics on term occurrences. Thus, this research can be fairly valuable in that it shows the direct correlation between term occurrence and syntactic functions of an NP. Moreover, the results are incorporated into an ATE system to evaluate the term status of term candidates. Different from previous methods, the proposed SF-Value is purely linguistically grounded with a focus on syntactic properties of NPs. Though, after combing it with C-Value, this ATE system do not have much better performance in terms of overall performance, it makes efforts to study whether syntactic functions can be used effectively to help select terms.

What's more, during the linguistic processing of these corpora, it is found that different parsers present linguistic information of different granularities, which will subsequently dictate the accuracy of term recognition. A parser which can produce more detailed syntactic analysis will influence the performance of ATE to a great extent.

References

Ananiadou S. (1994). A methodology for automatic term recognition. In *Proceedings of COLING 94*, pp. 1034-1038.

Barrón-Cedeño A., Sierra G., Drouin P. and Ananiadou S. (2009). An Improved Automatic Term Recognition Method for Spanish. In *Computational Linguistics and Intelligent Text Processing*, vol. 5449, Berlin-Heidelberg: Springer, pp. 125-136.

Bourigault D. (1992). Surface grammatical analysis for the extraction of terminological noun phrases. In *Proceedings of the 14th conference on Computational linguistics-Volume 3*, pp. 977-981.

Church K. and Hanks P. (1989). Word Association Norms, Mutual Information and Lexicography. *Computational Linguistics*, vol. (16:1): 22-29.

Church K., Gale W., Hanks P., Kindle D. and Bell Laboratories and Oxford University Press (1991). Using statistics in lexical analysis. In Zernik, U., editor, *Lexical Acquisition: Exploiting On-Line Resources to Build a Lexicon*, Hillsdale (New Jersey): Erlbaum, pp. 115-164.

Dagan, I. and Church K. (1994). Termight: identifying and translating technical terminology. In *Proceedings of the fourth conference on Applied natural language processing*, pp. 34-40.

Drouin P. (2006). Termhood experiments: quantifying the relevance of candidate terms. In Pitch, H., editor, *Modern Approaches to Terminological Theories and Applications, Linguistic Insights*, vol. 36, pp. 375-391.

Dunning T. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19(1): 61-74.

Eumeridou E. Nkwenti-Azeh B. and Mcnaught J. (2002). The contribution of verbal semantic content towards term recognition. *International Journal of Corpus Linguistics*, 7, 1: 87-106.

Eumeridou E., Nkwenti-Azeh B. and Mcnaught J. (2004). An Analysis of Verb Subcategorization Frames in Three Special Language Corpora with View towards Automatic Term Recognition. *Computers and the Humanities*, 38: 37-60.

Fang A.C. (1996). The Survey Parser: Design and Development. In Greenbaum, S., editor, *Comparing*

English World Wide: The International Corpus of English, Oxford: Oxford University Press, pp. 142-160.

Fang A.C. (2000). Evaluating the Performance of the Survey Parser with the NIST Scheme. In *CICLing 2006*, pp. 168-179.

Frantzi K., Ananiadou S. and Tsujii J. (1998). The C-value/NC-value Method of Automatic Recognition of Multi-word Terms. In *ECDL*, pp. 585-604.

Gillam L., Tariq M. and Ahmad K. (2005). Terminology and the Construction of Ontology. *Terminology*, 11(1): 55-81.

Jacquemin C. (1996). What is the tree that we see through the window: A linguistic approach to windowing and termvariation. *Information Processing and Management*, 32(4): 445-458.

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

Kageura K. and Umino B. (1996). Methods of Automatic Term Recognition: A review. *Terminology*, 3 (2): 259-289.

Kit C. (2002). Corpus Tools for Retrieving and Deriving Termhood Evidence. In *5th East Asia Forum of Terminology. Haikou: East Asia Forum on Terminology*, pp. 69-80.

Maynard D. and Ananiadou S. (2000). TRUCKS: A model for automatic multi-word term recognition. *Journal of Natural Language Processing*, 8 (1): 101-125.

Medical Subject Headings: http://www.nlm.nih.gov/pubs/factsheets/mesh.html.

Mima H. and Anaiadou S. (2001). An application of evaluation of the C/NC-value approach for the automatic term recognition of multi-word units in Japanese. *International Journal of Terminology*, 62, 2: 175-194.

Salton G. and Buckley C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing and Management: an International Journal*, 24 (5): 513-523.

Sclano F. and Velardi P. TermExtractor: A Web application to learn the common terminology of interest groups and research communities. In *Proceedings of the 9th Conf. on Terminology and Artificial Intelligence (TIA 2007)*, Sophia Antinopolis, France.

Smadja F., McKeown K.R. and Hatzivassiloglou V. (1996). Translating collocations for bilingual lexicons: a statistical approach. *Computational Linguistics*, 22 (1): 1-38.

VU Thuy, Aiti Aw and Min Zhang (2008) Term extraction through unithood and termhood unification. In *IJCNLP*, Jan 2008.

Wermter, J. and Hahn U. (2005). Effective Grading of Termhood in Biomedical Literature. In *AMIA Annual Symposium Proceedings*, 809.