Combining C-value and Keyword Extraction Methods for Biomedical Terms Extraction

Juan Antonio Lossio Ventura, Clement Jonquet
LIRMM, CNRS, Univ. Montpellier 2
Montpellier, France
fName.lName@lirmm.fr

Mathieu Roche, Maguelonne Teisseire
TETIS, Cirad, Irstea, AgroParisTech
Montpellier, France
fName.lName@teledection.fr

Abstract

The objective of this work is to extract and to rank biomedical terms from free text. We present new extraction methods that use linguistic patterns specialized for the biomedical field, and use term extraction measures, such as *C-value*, and keyword extraction measures, such as *Okapi BM25*, and *TFIDF*. We propose several combinations of these measures to improve the extraction and ranking process. Our experiments show that an appropriate harmonic mean of *C-value* used with keyword extraction measures offers better precision results than used alone, either for the extraction of single-word and multi-words terms. We illustrate our results on the extraction of English and French biomedical terms from a corpus of laboratory tests. The results are validated by using UMLS (in English) and only MeSH (in French) as reference dictionary.

1 Introduction

Language evolves faster than our ability to formalize and catalog concepts or possible alternative terms of these concepts. This is even more true for French in which the number of terms formalized in terminologies is significantly less important than in English. That is why our motivation is to improve the precision of automatic terms extraction process. Automatic Term Recognition (ATR) is a field in language technology that involves the extraction of technical terms from domain-specific language corpora (Zhang et al., 2008). Similarly, Automatic Keyword Extraction (AKE) is the process of extracting the most relevant words or phrases in a document with the propose of automatic indexing. Keywords, which we define as a sequence of one or more words, provide a compact representation of a document's content; two popular AKE measures are *Okapi BM25* (Robertson et al., 1999) and *TFIDF* (also called weighting measures). These two fields are summarized in Table 1.

	ATR	**AKE**
Input	one large corpus	single document
Output	terms of a domain	keywords of a doc
Domain	very specific	none
Exemples	*C-value*	*TFIDF*, *Okapi*

Table 1: Differences between ATR and AKE.

In our work, we adopt as baselines an ATR method, *C-value* (Frantzi et al., 2000), and the best two AKE methods (Hussey et al., 2012), previously mentioned and considered state-of-the-art. Indeed, the C-value, compared to other ATR methods, often gets best precision results and specially in biomedical studies (Knoth et al., 2009), (Zhang et al., 2008), (Zhang et al., 2004). Moreover, *C-value* is defined for multi-word term extraction but can be easily adapted for single-word term and it has never been applied to French biomedical text, which is appealing in our case.

Our experiments present a great improvement of the precision with these new combined methods. We give priority to precision in order to focus on extraction of new valid terms (i.e., for a candidate term to be a valid biomedical term or not) rather than on missed terms (recall).

The rest of the paper is organized as follows: Section 2 describes the related work in the field of ATR, and specially the uses of the *C-value*; Section 3 presents our combination of measures for ranking candidate terms; Section 4 shows and discusses our experiment results; and Section 5 concludes the paper.

2 Related work

ATR studies can be divided into four main categories: (i) rule-based approaches, (ii) dictionary-based approaches, (iii) statistical approaches, and (iv) hybrid approaches. Rule-based approaches for

instance (Gaizauskas et al., 2000), attempt to recover terms thanks to the formation patterns, the main idea is to build rules in order to describe naming structures for different classes using orthographic, lexical, or morphosyntactic characteristics. Dictionary-based approaches use existing terminology resources in order to locate term occurrences in texts (Krauthammer et al., 2004). Statistical approaches are often built for extracting general terms (Eck et al., 2010); the most basic measure is *frequency*. *C/NC-value* (Frantzi et al., 2000), is another statistical method well known in the literature that combines statistical and linguistic information for the extraction of multi-word and nested terms. While most studies address specific types of entities, *C/NC-value* is a domain-independent method. It was also used for recognizing terms from biomedical literature (Hliaoutakis et al., 2009). The *C/NC-value* method was also applied to many different languages besides English (Frantzi et al., 2000) such as Japanese (Mima et al., 2001), Serbian (Nenadić et al., 2003), Slovenian (Vintar, 2004), Polish (Kupsc, 2006), Chinese (Ji et al., 2007), Spanish (Barrón et al., 2009), and Arabic (Khatib et al., 2010), however to the best of our knowledge not to French. An objective of this work is to combine this method with AKE methods and to apply the combined measures to English and French. We believe that the combination of biomedical term extraction and the extraction of keywords describing a document, could be beneficial since keywords techniques give greater importance to the actual terms of this domain. This combination has never been proposed and experimented in the literature.

3 Proposed Methodology for Automatic Biomedical Term Extraction

This section describes the baselines measures and their customizations as well as the new combinations of these measures that we propose for automatic biomedical terms extraction and ranking. Our method for automatic term extraction has four main steps: (1) Part-of-Speech tagging, (2) Candidate terms extraction,(3) Ranking of candidate terms, (4) Computing of new combined measures.

Note, *C-value* is a method that deals with an unique corpus as input whereas AKE methods deal with several documents (cf. Table 1) then we need to do the union of documents for *C-value* to consider the whole corpus as an unique document. A preliminary step is the creation of patterns for French and English, as described hereafter.

3.1 Part-of-Speech tagging

Part-of-speech (POS) tagging is the process of assigning each word in a text to its grammatical category (e.g., noun, adjective). This process is performed based on the definition of the word or on the context which it appears in.

We apply part-of-speech to the whole corpus. We evaluated three tools (TreeTagger, Stanford Tagger and Brill's rules), and finally choose TreeTagger which gave best results and is usable both for French and English.

3.2 Candidate terms extraction

As previously cited work, we supposed that biomedical terms have similar syntactic structure. Therefore, we build a list of the most common lexical patterns according the syntactic structure of biomedical terms present in the UMLS[1] (for English) and the French version of MeSH[2] (for French). We also do a part-of-speech tagging of the biomedical terms using TreeTagger[3], then compute the frequency of syntactic structures. We finally choose the 200 highest frequencies to build the list of patterns for each language. The number of terms used to build these lists of patterns was 2 300 000 for English and 65 000 for French.

Before applying measures we filter out the content of our input corpus using patterns previously computed. We select only the candidate terms which syntactic structure is in the patterns list.

3.3 Ranking of candidate terms

3.3.1 Using *C-value*

The *C-value* method combines linguistic and statistical information (Frantzi et al., 2000); the linguistic information is the use of a general regular expression as linguistic patterns, and the statistical information is the value assigned with the *C-value* measure based on frequency of terms to compute the *termhood* (i.e., the association strength of a term to domain concepts). The aim of the *C-value* method is to improve the extraction of nested terms, it was specially built for extracting multi-word terms.

$$C\text{-}value(a) = \begin{cases} w(a) \times f(a) & \text{if } a \notin nested \\ w(a) \times \left(f(a) - \frac{1}{|S_a|} \times \sum_{b \in S_a} f(b) \right) & \text{otherwise} \end{cases} \quad (1)$$

[1] http://www.nlm.nih.gov/research/umls

[2] http://mesh.inserm.fr/mesh/

[3] www.cis.uni-muenchen.de/~schmid/tools/TreeTagger

Where a is the candidate term, $w(a) = \log_2(|a|)$, $|a|$ the number of words in a, $f(a)$ the frequency of a in the unique document, S_a the set of terms that contain a and $|S_a|$ the number of terms in S_a. In a nutshell, *C-value* either uses frequency of the term if the term is not include in other terms (first line), or decrease this frequency if the term appears in other terms, by using the frequency of those other terms (second line).

We modified the measure in order to extract all terms (single-word + multi-words terms), as suggested in (Barrón et al., 2009) in different manners: in the formula $w(a) = \log_2(|a|)$, we use $w(a) = \log_2(|a| + 1)$ in order to avoid null values (for single-word terms). Note that we do not use a stop word list nor a threshold for frequency as it was originally proposed.

3.3.2 Using *Okapi - TFIDF*

Those measures are used to associate each term of a document with a weight that represents its relevance to the meaning of the document it appears relatively to the corpus it is included in. The output is a ranked list of terms for each document, which is often used in information retrieval, to order documents by their importance given a query (Robertson et al., 1999). *Okapi* can be seen as an improvement of *TFIDF* measure, taking into account the document length.

The outputs of *Okapi* and *TFIDF* are calculated with a variable number of data so their values are heterogeneous. To manipulate these lists, the weights obtained from each document must be normalized. Once values normalized we have to merge the terms into a single list unique for the whole corpus to compare the results. Clearly the precision will depend on the method used to perform such merging. We merged following three functions, which calculate respectively the sum(S), max(M) and average(A) of the measures values of the term in whole the corpus. At the end of this task we have three lists from *Okapi* and three lists from *TFIDF*. The notation for these lists are $Okapi_X(a)$ and $TFIDF_X(a)$, where a is the term, X the factor $\in \{M, S, A\}$. For example, $Okapi_M(a)$ is the value obtained by taking the maximum Okapi value for a term a in the whole corpus.

3.4 Computing the New Combined Measures

With the goal of improving the precision of terms extraction we have conceived two new combined measures schemes, described hereafter, taking into account the values obtained in the above steps.

3.4.1 F-OCapi and F-TFIDF-C

Considered as the harmonic mean of the two used values, this method has the advantage of using all the values of the distribution.

$$F\text{-}OCapi_X(a) = 2 \times \frac{Okapi_X(a) \times C\text{-}value(a)}{Okapi_X(a) + C\text{-}value(a)} \quad (2)$$

$$F\text{-}TFIDF\text{-}C_X(a) = 2 \times \frac{TFIDF_X(a) \times C\text{-}value(a)}{TFIDF_X(a) + C\text{-}value(a)} \quad (3)$$

3.4.2 C-Okapi and C-TFIDF

Our assumption is that *C-value* can be more representative if the frequency, in Equation (1), of the terms is replaced with a more significant value, in this case the *Okapi's* or *TFIDF's* values of the terms (over the whole corpus).

$$C\text{-}m_X(a) = \begin{cases} w(a) \times m_X(a) & \text{if } a \notin nested \\ w(a) \times \left(m_X(a) - \frac{1}{|S_a|} \times \sum_{b \in S_a} m_X(b) \right) & \\ \text{otherwise} & \end{cases}$$

Where $m_X(a) = \{Okapi_X | TFIDF_X\}$, and $X \in \{M, S, A\}$.

4 Experiments and Results

4.1 Data and Experimental Protocol

We used biological laboratory tests, Labtestonline.org, as *corpus*. This site provides information in several languages to patient or family caregiver on clinical lab tests. Each test which forms a document in our corpus, includes the *formal lab test name*, some *synonyms* and possible *alternate names* as well as a description of the test. Our extracted corpus contains 235 clinical tests (about 400 000 words) for English and 137 (about 210 000 words) for French.

To automatically validate our candidate terms we compute a validation dictionary that include the *official name*, the *synonyms* and *alternate names* of the labtestonline tests plus all UMLS terms for English and the MeSH terms for French. These terminologies are references in the domain therefore each extracted term found in those is validated as a true term. Note that as a consequence we obtain 100% *Recall* with the whole list of extracted terms.

4.2 Experiments and results

Results are evaluated in terms of *Precision* obtained over the top k terms at different steps of our work presented in previous section. $Okapi$ and $TFIDF$ provided three lists of ranked candidate terms (M, S, A). For each combined measure using $Okapi$ or $TFIDF$, the experiments are done with the three lists. Therefore, the number of ranked list to compare is $C\text{-}value(1) + Okapi(3) + TFIDF(3) + F\text{-}OCapi(3) + F\text{-}TFIDF\text{-}C(3) + C\text{-}Okapi(3) + C\text{-}TFIDF(3) = 19$. In addition we experimented either for all (single and multi) or multi terms which finally give 38 ranked lists. Then, we select all terms (single and multi) or only muli-terms ($19 \times 2 = 38$ experiments for each language).

The following lines show part of the experiment results done all or multi terms, only and considering the top 60, 300 and 900 extracted terms, because it is appropriate and easier for an expert to evaluate the first best extracted terms. Table 2 and Table 3 compare the precision between the best baselines measures and the best combined measures. Best results were obtained in general with $F\text{-}TFIDF\text{-}C_M$ for English and $F\text{-}OCapi_M$ for French. **These tables prove that the combined measures based on the harmonic mean are better than the baselines measures, and specially for multi word terms, for which the gain in precision reaches 16%**. This result is particularly positive because in the biomedical domain it is often more interesting to extract multi-word terms than single-word terms. However, one can notice that results obtained to extract all terms with $C\text{-}Okapi_S$ and $C\text{-}TFIDF_S$ are not better than $Okapi_X$ or $TFIDF_X$ used directly. The reason is because the performance of those new combined measures is affected when single word terms are extracted. Definitively, the new combined measures are really performing for multi word term.

Results of AKE methods for English show that $TFIDF_X$ obtains better results than $Okapi_X$. The main reason for this, is because the size of the English corpus is larger than the French one, and *Okapi* is known to perform better when the corpus size is smaller (Lv et al., 2011).

In addition, Table 3 shows that *C-value* can be used to extract French biomedical terms with a better precision than what has been obtained in previous cited works with different languages. The precision of *C-value* for the previous work was between 26% and 31%.

	All Terms			Multi Terms		
	60	300	900	60	300	900
$Okapi_M$	0.96	0.95	0.82	0.68	0.62	0.54
$Okapi_S$	0.83	0.89	0.85	0.58	0.57	0.55
$Okapi_A$	0.72	0.31	0.27	0.48	0.39	0.26
$TFIDF_M$	0.97	0.96	0.84	0.71	0.63	0.54
$TFIDF_S$	0.96	0.95	**0.93**	0.82	0.71	0.61
$TFIDF_A$	0.78	0.74	0.63	0.50	0.40	0.37
$C\text{-}value$	0.88	0.92	0.89	0.72	0.71	0.62
$F\text{-}OCapi_M$	0.73	0.87	0.84	0.79	0.69	0.58
$F\text{-}TFIDF\text{-}C_M$	**0.98**	**0.97**	0.86	**0.98**	**0.73**	**0.65**
$C\text{-}Okapi_S$	0.88	0.86	0.80	0.61	0.58	0.53
$C\text{-}TFIDF_S$	0.96	0.95	0.86	0.85	0.71	0.61

Table 2: Extract of precision comparison for term extraction for English.

	All Terms			Multi Terms		
	60	300	900	60	300	900
$Okapi_M$	**0.90**	0.61	0.37	0.53	0.31	0.18
$Okapi_S$	0.30	0.31	0.37	0.23	0.30	0.37
$Okapi_A$	0.52	0.31	0.16	0.30	0.17	0.16
$TFIDF_M$	0.75	0.51	0.37	0.45	0.28	0.18
$TFIDF_S$	0.68	0.48	0.42	0.53	0.33	0.22
$TFIDF_A$	0.12	0.39	0.29	0.17	0.16	0.11
$C\text{-}value$	0.43	0.42	0.43	0.35	0.34	**0.26**
$F\text{-}OCapi_M$	0.73	**0.62**	**0.43**	**0.65**	**0.35**	0.22
$F\text{-}TFIDF\text{-}C_M$	0.85	0.57	0.39	0.62	0.31	0.19
$C\text{-}Okapi_S$	0.28	0.32	0.34	0.23	0.28	0.20
$C\text{-}TFIDF_S$	0.65	0.55	0.38	0.50	0.32	0.19

Table 3: Extract of precision comparison for term extraction for French.

We also have done experiments with two more corpus: (i) the Drugs data from MedlinePlus[4] in English and, (ii) PubMed[5] citations' titles in English and French, we have verified that the new combined measures are performing better, particularly these based on the harmonic mean, $F\text{-}TFIDF\text{-}C_M$ and $F\text{-}OCapi_M$.

5 Conclusions and Perspectives

This work present a methodology for term extraction and ranking for two languages, French and English. We have adapted *C-value* to extract French biomedical terms, which was not proposed in the literature before. We presented and evaluated two new measures thanks to the combination of three existing methods. The best results were obtained by combining *C-value* with the best results from AKE methods, i.e., $F\text{-}TFIDF\text{-}C_M$ for English and $F\text{-}OCapi_M$ for French.

For our future evaluations, we will enrich our dictionaries with BioPortal's[6] terms for English and CISMeF's[7] terms for French. Our next task will be the extraction of relations between these new terms and already known terms, to help in ontology population. In addition, we are currently implementing a web application that implements these measures for the community.

[4] http://www.nlm.nih.gov/medlineplus/

[5] http://www.ncbi.nlm.nih.gov/pubmed

[6] http://bioportal.bioontology.org/

[7] http://www.chu-rouen.fr/cismef/

Acknowledgments

This work was supported in part by the French National Research Agency under JCJC program, grant ANR-12-JS02-01001, as well as by University Montpellier 2 and CNRS.

References

Alberto Barrón-Cedeño, Gerardo Sierra, Patrick Drouin, Sophia Ananiadou. 2009. An Improved Automatic Term Recognition Method for Spanish. *Proceeding of Computational Linguistics and Intelligent Text Processing,* pp 125-136.

NeesJan Eck, Ludo Waltman, EdC.M. Noyons, ReindertK Buter. 2010. Automatic term identification for bibliometric mapping. *SpringerLink, Scientometrics,* Volume 82, Number 3.

Katerina Frantzi, Sophia Ananiadou, Hideki Mima 2000. Automatic Recognition of Multi-Word Terms: the C-value/NC-value Method. *International Journal of Digital Libraries, 3(2)* pp.117-132.

Robert Gaizauskas, George Demetriou, Kevin Humphreys. 2000. Term Recognition and Classification in Biological Science Journal Articles. *Proceedings of the Computational Terminology for Medical and Biological Applications Workshop*, pp 37–44.

Angelos Hliaoutakis, Kaliope Zervanou and Euripides G.M. Petrakis. 2009. The AMTEx approach in the medical document indexing and retrieval application. *Data and Knowledge Eng.,* pp 380-392.

Richard Hussey, Shirley Williams, Richard Mitchell. 2012. Automatic keyphrase extraction: a comparison of methods. *Proceedings of the International Conference on Information Process, and Knowledge Management,* pp. 18-23.

Luning Ji, Mantai Sum, Qin Lu, Wenjie Li, Yirong Chen. 2007. Chinese Terminology Extraction Using Window-Based Contextual Information. *Proceeding of CICLing, LNCS,* pp.62–74.

Khalid Al Khatib, Amer Badarneh. 2010. Automatic extraction of Arabic multi-word terms. *Proceeding of Computer Science and Information Technology.* pp 411-418.

Petr Knoth, Marek Schmidt, Pavel Smrz, Zdenek Zdrahal. 2009. Towards a Framework for Comparing Automatic Term Recognition Methods. *Conference Znalosti*.

Michael Krauthammer, Goran Nenadic. 2004. Term identification in the biomedical literature. *Journal of Biomedical Informatics,* pp 512–526.

Anna Kupsc. 2006. Extraction automatique de termes à partir de textes polonais. *Journal Linguistique de Corpus.*

Yuanhua Lv, ChengXiang Zhai. 2011. When documents are very long, BM25 fails! *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval,* pp.1103–1104.

Olena Medelyan, Eibe Frank, Ian H. Witten. 2009. Human-competitive tagging using automatic keyphrase extraction. *Proceeding of the International Conference of Empirical Methods in Natural Language Processing, Singapore*.

Hideki Mima, Sophia Ananiadou, . 2001. An application and evaluation of the C/NC-value approach for the automatic term recognition of multi-word units in Japanese. *Japanese Term Extraction. Special issue of Terminology,* vol 6:2

Goran Nenadić, Irena Spasić, Sophia Ananiadou. 2003. Morpho-syntactic clues for terminological processing in Serbian. *Proceeding of the EACL Workshop on Morphological Processing of Slavic Languages,* pp.79–86.

Natalya F. Noy, Nigam H. Shah, Patricia L. Whetzel, Benjamin Dai, Michael Dorf, Nicholas B. Griffith, Clement Jonquet, Daniel L. Rubin, Margaret-Anne Storey, Christopher G. Chute, Mark A. Musen. 2009. BioPortal: ontologies and integrated data resources at the click of a mouse. *Nucleic Acids Research,* pp. 170-173 vol. 37.

Stephen Robertson, Steve Walker, Micheline Hancock-Beaulieu. 1999. Okapi at TREC-7: automatic ad hoc, filtering, VLC and interactive track. *IN*. pp. 253–264 vol. 21.

Francesco Sclano, Paola Velardi. 2007. TermExtractor: a Web Application to Learn the Common Terminology of Interest Groups and Research Communities. *In Enterprise Interoperability II,* pp. 287-290.

Špela Vintar. 2004. Comparative Evaluation of C-Value in the Treatment of Nested Terms. *Workshop (Methodologies and Evaluation of Multiword Units in Real-world Applications),* pp.54–57.

Yongzheng Zhang, Evangelos Milios, Nur Zincirheywood. 2004. A Comparison of Keyword- and Keyterm-Based Methods for Automatic Web Site Summarization. *AAAI04 Workshop on Adaptive Text Extraction and Mining,* pp. 15–20.

Ziqi Zhang, José Iria, Christopher Brewster, Fabio Ciravegna. 2008. A Comparative Evaluation of Term Recognition Algorithms. *Proceedings of the Sixth International Conference on Language Resources and Evaluation*.