Towards a mixed approach to extract biomedical terms from text corpus

Juan Antonio Lossio-Ventura, Clement Jonquet, Mathieu Roche, Maguelonne Teisseire

Towards a Mixed Approach to Extract Biomedical Terms from Text Corpus

Juan Antonio Lossio Ventura, LIRMM, University Montpellier 2, Montpellier, France & CNRS, Paris, France

Clement Jonquet, LIRMM, University Montpellier 2, Montpellier, France & CNRS, Paris, France

Mathieu Roche, UMR TETIS, Cirad, Irstea, AgroParisTech, Montpellier, France

Maguelonne Teisseire, UMR TETIS, Cirad, Irstea, AgroParisTech, Montpellier, France

ABSTRACT

The objective of this paper is to present a methodology to extract and rank automatically biomedical terms from free text. The authors present new extraction methods taking into account linguistic patterns specialized for the biomedical domain, statistic term extraction measures such as C-value and statistic keyword extraction measures such as Okapi BM25, and TFIDF. These measures are combined in order to improve the extraction process and the authors investigate which combinations are the more relevant associated to different contexts. Experimental results show that an appropriate harmonic mean of C-value associated to keyword extraction measures offers better precision, both for single-word and multi-words term extraction. Experiments describe the extraction of English and French biomedical terms from a corpus of laboratory tests available online. The results are validated by using UMLS (in English) and only MeSH (in French) as reference dictionary.

Keywords: *Biomedical Natural Language Processing (BioNLP), Biomedical Term Extraction, Biomedical Terminologies and Ontologies, Biomedical Thesaurus, Statistic Measure, Text Mining*

INTRODUCTION

The huge amount of data available online today is often composed of plain text field, for instances, clinical trial descriptions, adverse event reports or electronic health records. These texts often contain the real language (expressions and terms) used by the community. Although in the biomedical domain there exist hundred of terminologies and ontologies to describe such languages (Noy et al., 2009), those terminologies often miss concepts or possible alternative terms for those concepts. Our motivation is to improve the precision of automatic terms extraction process, the main reason for this, is that language evolves faster than our ability to formalize and catalog it. This is even more true for French in which the number of terms

DOI: 10.4018/ijkdb.2014010101

formalized in terminologies is significantly less important than in English.

NLP (natural language processing) tools and methods enable to enrich biomedical dictionaries from texts. Automatic Term Recognition (ATR) is an approach in language technology that involves the extraction of technical terms from domain-specific language corpora (Zhang et al., 2008). In addition, Automatic Keyword Extraction (AKE) is the process of extracting the most relevant words or phrases in a document. 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* and *TFIDF*, also called weighting measures. These two fields are summarized in *Table 1*.

In our work, we adopt as baseline measures an ATR method, *C-value* (Frantzi et al., 2000), and the best two AKE methods (Hussey et al., 2012). Indeed, the *C-value*, compared to other ATR methods, often gets best precision results and especially in biomedical studies (Knoth et al., 2009; Zhang et al., 2008; Zhang et al., 2004). Moreover, this measure is defined for multi-word term extraction but can be easily adapted for single-word term (presented later on) and it has never been applied to French text, which is appealing in our case. *Okapi* and *TFIDF* are the best AKE methods (Hussey et al., 2012). We propose to define new extraction methods by combining in different manners ATR and AKE measures, in order to rank the best candidate terms. Our experiment results underline the precision efficiency gain with the proposed methods. We give priority to precision in order to focus on extraction of new valid terms (precision) rather than on missed terms (recall), i.e., for a candidate term to be a valid biomedical term or not.

The rest of the paper is organized as follows: section "Related Work" describes the state of the art in the field of ATR, and specially the methods based on C-value; section "Proposed Approach" presents our proposal of ranking measures; section "Experiments and Results" details and discusses the conducted experiments and the associated results; and section "Conclusion" concludes the paper.

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, used for extracting terms

Table 1. Differences between ATR and AKE

	Automatic Term Recognition (ATR)	Automatic Keyword Extraction (AKE)
Input	one large corpus (i.e., not explicitly separated in documents)	single document within a dataset of documents
Output	technical terms of a domain	keywords that describe the document
Domain	very specific	none
Exemples	*C-value*	*TFIDF, Okapi*

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, Serbian (Nenadic´ 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). To the best of our knowledge, it has never been used to French texts.

The main objective of our work is thus to combine this method with AKE methods and to evaluate them both for English and French. Indeed, we argue that the combination of biomedical term extraction and keywords extraction methods could highlight relevant terms of biomedical domain.

PROPOSED APPROACH

This section describes the baseline measures and their customizations as well as new combinations of these measures for automatic biomedical terms extraction. In subsection A, we detail the extensions of the baselines measures. Particularly, we improve the *C-value* method by taking into consideration linguistic pattern specialized for biomedical domain. In addition, we adapt the statistic measure in order to extract single and multi terms. These approaches are applied both to French and English languages. We also use *Okapi BM25* (hereafter *Okapi*) and *TFIDF*. Subsection B presents some proposed combinations of the basic measures: (i) Computing harmonic mean combinations, (ii) Taking into account the *Okapi* values and *TFIDF* values within the calculus of *C-value*.

Our method for automatic term extraction has four main steps (Lossio et al., 2013), described in *Figure 1*:

1. Part of Speech tagging of the corpus,
2. Candidate terms extraction following patterns,
3. Ranking of candidate terms,
4. Computing new combined measures.

We execute those four steps by taking either *C-value* (right branch) or *Okapi*/*TFIDF* (left branch) as baseline methods. Notice that as the input of *C-value* is a unique element and the weighting measure deals with many documents (cf. Table 1), we need to merge all documents to build a single textual element. A preliminary step not represented in Figure 1 is the creation of patterns for French and English, as described hereafter.

Building Biomedical Patterns

We consider the following assumption: biomedical terms have similar syntactic structure. Therefore, we build a list of the most common lexical patterns according the syntactic structure of terms that are in biomedical databases, UMLS[1] (Unified Medical Language System) for English and MeSH[2] (Medical Subject Headings) for French.

First, a part-of-speech tagging of the biomedical terms is done by using TreeTagger.[3] The frequency of syntactic structures is then computed. The top-200 are selected as patterns for each language. The number of terms used to build the list was 2 300 000 for English and 65 000 for French. Examples of patterns, sorted by frequency, are given in Table 2.

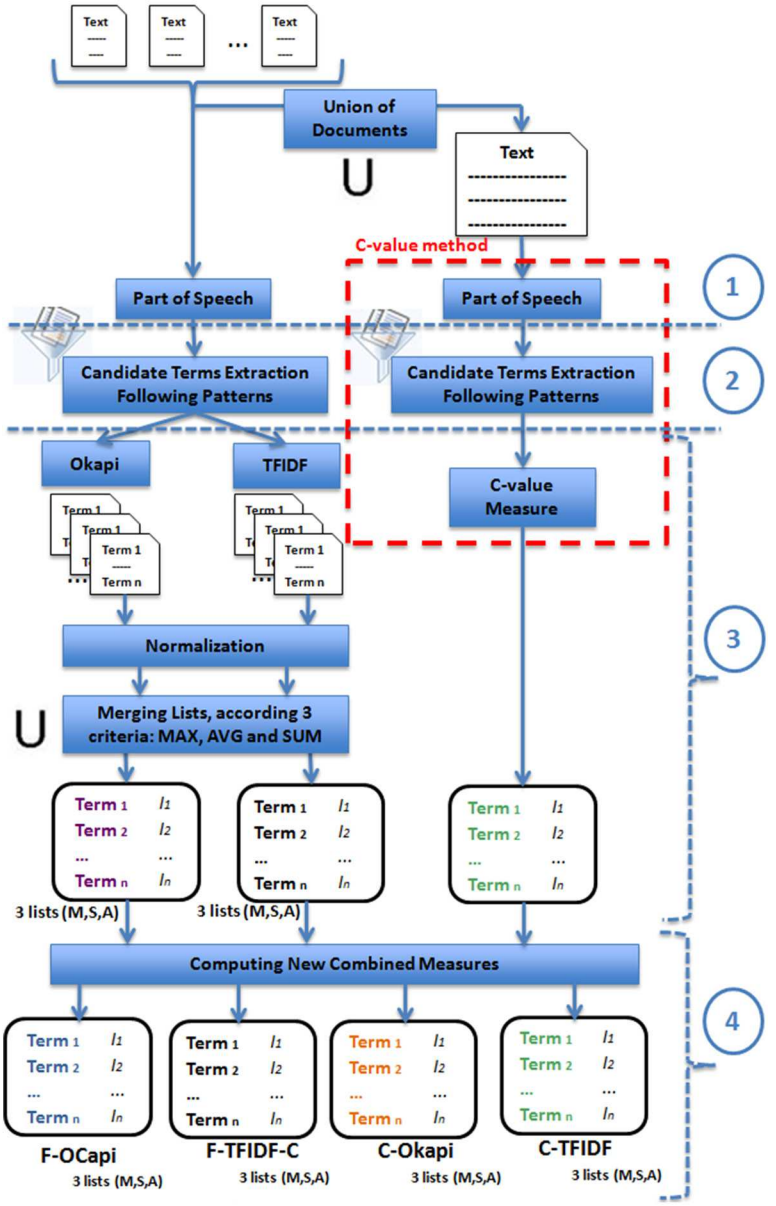
Part-of-Speech Tagging (See Part (1) in Figure 1)

Part-of-speech (POS) tagging assigns each word in a text to its grammatical category (e.g., noun, adjective). This process is based on the definition of the word or on the context in which it appears. At this step, as suggested in the *C-value* method, the part-of-speech is applied on the whole corpus. We evaluated three tools (TreeTagger, Stanford Tagger and Brill's rules) and finally choose TreeTagger which gave better results and is usable both for French and English.

Candidate Terms Extraction Based on Biomedical Patterns (see Part (2) in Figure 1)

Before applying any measures we filter out the content of our input corpus using patterns previously computed. We select only the terms

Figure 1. Workflow of biomedical term extraction



which syntactic structure is in the patterns list. Of course, the pattern filtering occurs specifically by language (i.e., when text is in French, only French list of patterns is used).

- **Union Documents:** The *C-value* method needs a single text document as input. This step merges all texts of the corpus into one document.

Table 2. Examples of the 9 most frequent patterns for English and French

	English	French
1	ProperNoun	Noun
2	Noun	Noun Adj
3	ProperNoun ProperNoun	Noun Prep Noun
4	Noun Noun	Noun Adj Adj
5	Adj Noun	Noun Prep:det Noun
6	Noun Noun ProperNoun	Noun Prep ProperNoun
7	Adj ProperNoun ProperNoun	Noun ProperNoun
8	Noun ProperNoun ProperNoun	Noun Noun
9	Noun Noun Prep Noun	Noun Prep Noun Adj

Ranking of Candidate Terms (See Part (3) in Figure 1)

1. **Ranking Terms with *C-Value*:** The *C-value* method combines linguistic and statistical information (Frantzi et al., 2000). The linguistic information is based on the use of a general regular expression (i.e., linguistic patterns). The statistical information is the value assigned with the *C-value* measure based on the term frequency to compute the term hood (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 has been specially defined for extracting multi-word terms.

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

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 $w(a) = \log_2(|a| + \mathbf{1})$ in order to avoid null values (for single-word terms) as illustrated in Table 3. Note that we do not use a stop word list nor a threshold for frequency as it was originally proposed.

Table 3 shows the proposed changes for the computation of *w(a)* with the original and modified *C-value* definitions.

2. **Ranking Terms with *Okapi - TFIDF*:** In a nutshell, these measures are used to associate each occurrence of a term with a weight representing its relevance to the meaning of the document it appears in and relatively to the corpus it is included in (and also relatively to the size of the document in the case of *Okapi*). The output is a ranked list of terms for each document. They serve as ranking measures to order documents by their importance given a query (Robertson et al., 1999). *Okapi* can be seen as an improvement of the *TFIDF* measure, taking into account the document

Table 3. Calculation of w(a)

	Original *C-value*	**Modified** *C-value*
	w(a) = log2(|a|)	**w(a) = log2(|a| + 1)**
antiphospholipid antibodies	log2(2) = 1	log2(2 + 1) = 1.6
white blood	log2(2) = 1	log2(2 + 1) = 1.6
platelet	log2(1) = 0	log2(1 + 1) = 1

length. Both measures are mostly used for information retrieval and text mining.

a. **Normalization:** The *Okapi* and *TFIDF* measures are calculated with a variable number of elements, so that the obtained values are heterogeneous. In order to manipulate these result lists, the weights obtained from each document must be normalized for the whole corpus. Therefore, the results of each measure have to be normalized, for instance between 0 and 1.
b. **Merging Lists:** Once values normalized, we have to merge the terms into a single list in order to evaluate the results. Clearly, the precision will depend on the method used to perform such merging. We merged following three functions: Sum(S), Maximum(M), and Average(A) which calculate respectively the sum, max and average of a term in the whole collection. At the end of this task, we obtain 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 instance, $Okapi_M(a)$ is the list obtained by taking the maximum *Okapi* value for a term a in the whole corpus.

Computing the New Combined Measures (See Part (4) in Figure 1)

With aim of improving the precision of terms extraction, we have conceived two new combined measure schemes, taking into account the results obtained in the above steps. The first one is based on the harmonic mean of two values. The second one is obtained by replacing the *frequency*, within the Equation (1) of *C-value*, by the value of the weighting measures.

F-OCapi and F-TFIDF-C: Considered as the harmonic mean of the two used values, this method has as advantage to use all values of the distribution.

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

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

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

$$\text{C}-\boldsymbol{m}_x(\boldsymbol{a}) = \begin{cases} w(a)\times\boldsymbol{m}_x(a), & if\ a\notin nested \\ w(a)\times\left(\boldsymbol{m}_x(\boldsymbol{a}) - \frac{1}{|S_a|}\times\sum_{b\in S_a}\boldsymbol{m}_x(\boldsymbol{b})\right), & otherwise \end{cases} \quad (4)$$

Where $\boldsymbol{m}_{\boldsymbol{x}}\left(\boldsymbol{a}\right) = \left\{Okapi_X, TFIDF_X\right\}$, and $X \in \left\{M, S, A\right\}$.

Table 4 shows different ranking of terms with different measures. This example highlights specific and very relevant terms such as "antiphospholipid antibodies" and "platelet". Indeed these terms obtain a better ranking by using our measures such as *F-TFIDF-CM*.

Implementation and Availability

We developed *BioTex*, a web application (illustrated in Figure 2) that implements the entire workflow presented in this paper: for a given text corpus as input BioTex will extract and rank biomedical terms according to the selected extraction measure included in *C-value, Okapi*, *TFIDF* or one of the new proposed combinations. In addition, BioTex allows to validate automatically terms already existing in the available UMLS/MeSH-fr terminologies. We have implemented and evaluated BioTex for both English and French. The application is available online but can also be used in any program through a java API: http://tubo.lirmm.fr:8080/biotex.

In the following section, we evaluate a large list of extracted and ranked terms with our new measures and their different combinations (using our web application).

DATA AND EXPERIMENTAL PROTOCOL

Test Collection

We used biological laboratory tests as corpus, obtained from LabTestsOnline.org. This site provides information in several languages to patient or family caregiver on clinical lab tests. Each test, which is considered as a document in our corpus, includes the *formal lab test name*, its *synonyms* and many *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.

Validation Data

In order 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. We can now evaluate precision with a proper reference for valid terms. Note that as a consequence, the recall is equal to 100% with the whole list of extracted terms.

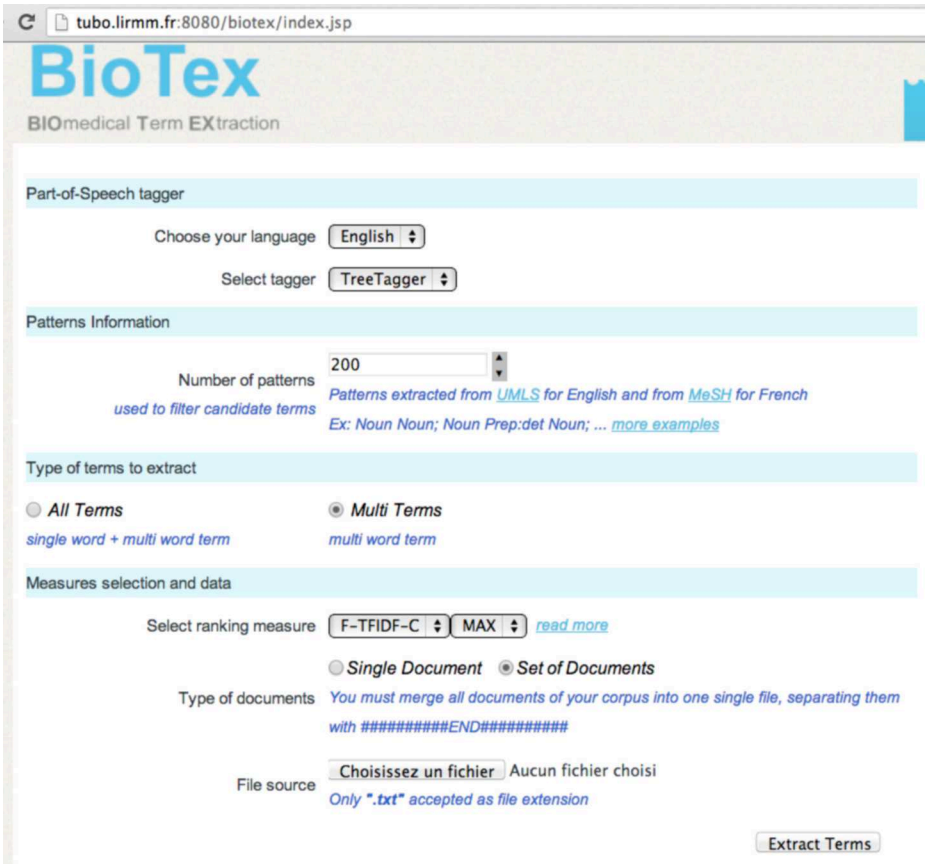
EXPERIMENTS AND RESULTS

A first evaluation was done automatically, without the verification of an expert to validate or invalidate the terms that are not found in our validation dictionary. Results are evaluated in terms of Precision obtained over the top k terms at different steps of our workflow presented in

Table 4. Rank of terms based on different measures

	Ranking of the terms						
	C-value	$TFIDF_M$	$Okapi_M$	*F-TFIDF-*C_M	*F-OCapi*$_M$	*C-TFIDF*$_M$	*C-Okapi*$_S$
antiphospholipid antibodies	**496**	**112**	162	**45**	141	8	1770
white blood	**129**	**745**	387	**796**	356	679	754
platelet	**159**	**112**	112	**15**	59	219	800

Figure 2. BioTex web application that implements the biomedical term extraction workflow



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 conducted with the three lists. Therefore, the number of ranked list to compare is 19: *C-value*(1) + *Okapi*(3) + *TFIDF*(3) + *F-OCapi*(3) + *F-TFIDF-C*(3) + *C-Okapi*(3) + *C-TFIDF*(3). In addition we experimented the workflow 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 sections 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 to evaluate only the *top-k* extracted terms. We evaluated first the baselines measures and second with the new combined measures for English and French.

Experiments with AKE Methods: Okapi and TFIDF

The experiments with these methods were performed after applying the linguistic filter. The experiments were carried for All and Multi terms extraction. Table 5 and Table 6 show the results of term extraction with $Okapi_X$. Best results were often obtained with $Okapi_M$ for both languages.

Table 7 and Table 8 show the results of terminology extraction with $TFIDF_X$. Best

Table 5. Precision of $Okapi_X$ on English corpus

	All Terms			Multi Terms		
	60 terms	**300 terms**	**900 terms**	**60 terms**	**300 terms**	**900 terms**
$Okapi_M$	**0,96**	**0,95**	0,82	**0,68**	**0,62**	**0,55**
$Okapi_S$	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

results were obtained with $TFIDF_M$ for All terms for both languages. For Multi terms, the best results were obtained with $TFIDF_S$, for both languages.

Experiments with C-value

In this subsection, we evaluated the ATR method, *C-value (see Tables 9 and 10)*.

Experiments with New Combined Measures

The new measures were also evaluated. Table 11 and Table 12 present the results of terminology extraction with these new measures. In general, the best precision rate is obtained with $F-TFIDF-C_M$ for English and $F-OCapi_M$ for French.

Manual Validation

In order to know the true precision, because in the manual validation there are terms that are not in our dictionaries. So, we export a list of extracted terms to be manually validated. For this, we choose the list with the best precision rate in the automatic validation process. Table 13 and Table 14 compare the best results of the above evaluated measures. In general, $F-TFIDF-C_M$ obtained the best results for English extraction terms and $F-OCapi_M$ obtains highest precision for biomedical French. Experts validated these two lists, composed of 300 terms. Table 15 and Table 16 show the precision computed with the manual validation compared to the one with the automatic validation. Note that the manual validation confirms that our ranking function has a good behavior because the precision value is better for first terms.

Table 6. Precision of $Okapi_X$ on French corpus

	All Terms			Multi Terms		
	60 terms	**300 terms**	**900 terms**	**60 terms**	**300 terms**	**900 terms**
$Okapi_M$	**0,90**	**0,61**	**0,37**	**0,53**	**0,31**	**0,37**
$Okapi_S$	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

Table 7. Precision of $TFIDF_X$ on English corpus

	All Terms			Multi Terms		
	60 terms	**300 terms**	**900 terms**	**60 terms**	**300 terms**	**900 terms**
$TFIDF_M$	**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

Table 8. Precision of $TFIDF_X$ on French corpus

	All Terms			Multi Terms		
	60 terms	**300 terms**	**900 terms**	**60 terms**	**300 terms**	**900 terms**
$TFIDF_M$	**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

DISCUSSION

In the results of AKE methods, *TFIDF* obtains better results than *Okapi*. 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).

Table 10 shows that C-value can be used to extract French biomedical terms with a

Table 9. Precision of C-value on English corpus

	All Terms			Multi Terms		
	60 terms	**300 terms**	**900 terms**	**60 terms**	**300 terms**	**900 terms**
C-value	0,88	0,92	**0,89**	0,72	**0,71**	**0,62**

Table 10. Precision of C-value on French corpus

	All Terms			Multi Terms		
	60 terms	**300 terms**	**900 terms**	**60 terms**	**300 terms**	**900 terms**
C-value	0,43	0,42	**0,43**	0,35	**0,35**	**0,26**

Table 11. Precision comparison of new measures for English

	All Terms			Multi Terms		
	60 terms	**300 terms**	**900 terms**	**60 terms**	**300 terms**	**900 terms**
$F-OCapi_M$	0,73	0,87	0,84	0,79	0,69	0,58
$F-TFIDF-C_M$	**0,98**	**0,97**	**0,86**	**0,98**	**0,73**	**0,65**
$C-Okapi_S$	0,88	0,86	0,80	0,61	0,58	0,53
$C-TFIDF_S$	0,96	0,95	**0,86**	0,85	0,71	0,61

better precision than what has been obtained in previous cited works in other languages. The precision of C-value for the previous works was between 26% and 31%.

For the new combined measures, the best results are obtained by combining *C-value* with the best results from AKE methods, i.e., *F-TFIDF-*C_M and *F-*$OCapi_M$. Table 13 and Table 14 compare the precision between the best baselines measures and the best combined measures. Best results were obtained in general with *F-TFIDF-*C_M for English and *F-*$OCapi_M$ for French. These figures 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 multiword terms than single-word terms. However, one can notice that results obtained to extract all terms with tcokapi and ctfidf are not better than okapi or tfidf use directly. The main reason for this is because the performance of those new combined measures are absorbed by the effect of extracting also single world terms. Definitively, all the new combined measures are really performing better for multi word terms.

Several terms proposed by our system are considered as irrelevant (i.e., false positive examples) with our automatic validation protocol because they are not present in known biomedical dictionaries, which does not mean that they are irrelevant. Actually elements that are not found in biomedical resources can be relevant thanks a manual validation. For

Table 12. Precision comparison of new measures for French

	All Terms			Multi Terms		
	60 terms	**300 terms**	**900 terms**	**60 terms**	**300 terms**	**900 terms**
$F-OCapi_M$	0,73	**0,62**	**0,43**	**0,65**	**0,35**	**0,22**
$F-TFIDF-C_M$	**0,85**	0,57	0,39	0,62	0,31	0,19
$C-Okapi_S$	0,28	0,32	0,34	0,23	0,28	0,20
$C-TFIDF_S$	0,65	0,55	0,38	0,50	0,32	0,19

Table 13. Precision of the best measures for the extraction of all terms for English

	All Terms			
	60 terms	**90 terms**	**300 terms**	**3000 terms**
$F-TFIDF-C_M$	**0,98**	**0,97**	0,86	**0,75**
$C-TFIDF_S$	0,96	0,95	0,86	0,68
C-value	0,88	0,92	**0,89**	0,73
$Okapi_M$	0,96	0,95	0,82	0,51
$TFIDF_M$	0,97	0,96	0,84	0,62

Table 14. Precision of the best measures for the extraction of all terms for French

	All Terms			
	60 terms	**90 terms**	**300 terms**	**3000 terms**
$F-OCapi_M$	0,73	**0,62**	**0,43**	**0,31**
$C-TFIDF_S$	0,65	0,55	0,38	0,22
C-value	0,43	0,42	0,43	0,29
$Okapi_M$	**0,90**	0,61	0,37	0,30
$TFIDF_M$	0,75	0,51	0,37	0,29

Table 15. Precision of $F-TFIDF-C_M$ for English with automatic and manual validations

	Multi Terms by $F-TFIDF-C_M$					
	30 terms	**60 terms**	**90 terms**	**120 terms**	**180 terms**	**300 terms**
Automatic Validation	96,67%	98,33%	87,78%	84,17%	77,78%	72,67%
Manual Validation	**100,00%**	**100,00%**	**99,17%**	**98,89%**	**96,67%**	**93,00%**

Table 16. Precision of $F-OCapi_M$ for French with automatic and manual validations

	Multi Terms by $F-OCapi_M$					
	30 terms	**60 terms**	**90 terms**	**120 terms**	**180 terms**	**300 terms**
Automatic Validation	63,33%	65,00%	53,33%	49,17%	39,44%	34,67%
Manual Validation	**100,00%**	**98,33%**	**95,56%**	**95,83%**	**95,00%**	**91,67%**

instance, they can represent new terms to add in biomedical dictionaries. So in Tables 15 and 16, the precision rate is naturally higher with a manual validation.

In addition to Labtestonline.org, we also have done experiments with two more corpus: (i) the Drugs data from MedlinePlus[4] which contains about 1.05 million of words in English; we have verified that the new combined measures are performing better, particularly these based on the harmonic mean, F-TFIDF-CM and F-OCapiM. (ii) PubMed[5] citations' titles in English and French, which contain about 2000 titles of articles; the results show a small difference between the baseline measures and the new combined measures mainly because titles are small piece of text and therefore the new combined measures cannot take advantage of the frequency.

CONCLUSION AND FUTURE WORK

This paper presents a new methodology to automatically extract biomedical terminology to propose relevant terms to experts. For term ranking, 19 measures have been proposed for two languages, French and English.

We have adapted *C-value* to extract French biomedical terms, which was not proposed in the literature before. The precision of the *C-value* in previous works was between 26% and 31%. With this proposal, we greatly improved these results. This measure has been improved by first adding linguistic patterns of biomedical field. Second, the statistical aspects of the measure have been changed in order to take into account all types of terms (i.e., single- and multi-word terms).

We applied two AKE methods, for extracting keywords from a document, merging the terms following three merging factors into a single list.

We presented and evaluated two new measures thanks to the combination of three existing methods. The evaluation showed that these combinations obtain the best precision rates for both cases, all and multi term extraction for French.

Future work will be dedicated to (i) a web ranking in order to improve the precision of the terminologies lists, and (ii) the improvement of the BioTex web application & web service inorder to enable anyone to query to use any of our proposed biomedical term extractions methods on other datasets. We are also considering to enrich our validation dictionaries with BioPortal[6] terms for English and CISMeF[7] terms for French.

ACKNOWLEDGMENT

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 and IBC Project (www.ibc-montpellier.fr).

REFERENCES

Al Khatib, K., & Amer Badarneh. (2010). Automatic extraction of Arabic multi-word terms. In *Proc. of Computer Science and Information Technology (IMCSIT)* (pp. 411-418).

Barrón-Cedeño, A., Sierra, G., Drouin, P., & Ananiadou, S. (2009). *An improved automatic term recognition method for Spanish* (pp. 125–136). Proc. of Computational Linguistics and Intelligent Text Processing. doi:10.1007/978-3-642-00382-0_10

Cilibrasi, R. L., & Vitanyi, P. M. B. (2007). The Google similarity distance. *IEEE Transactions on Knowledge and Data Engineering*, 370–383. doi:10.1109/TKDE.2007.48

Eck, N. J., Waltman, L., Noyons, E. C. M., & Buter, R. K. (2010). Automatic term identification for bibliometric mapping. *SpringerLink Scientometrics, 82*(3).

Frantzi, K., Ananiadou, S., & Mima, H. (2000). Automatic recognition of multi-word terms: The C-value/NC-value method. *International Journal on Digital Libraries*, *3*(2), 117–132. doi:10.1007/s007999900023

Gracia, J., Trillo, R., Espinoza, M., & Mena, E. (2006). Querying the web: A multiontology disambiguation method. In *Proceedings of the 6th International Conference on Web Engineering* (pp. 241–248).

Gracia, J., Trillo, R., Espinoza, M., & Mena, E. (2008). Web-based measure of semantic relatedness. In *Proceedings of the 9th International Conference on Web Information Systems Engineering* (pp. 136–150).

Hliaoutakis, A., Zervanou, K., & Petrakis, E. G. M. (2009). The AMTEx approach in the medical document indexing and retrieval application. *Data & Knowledge Engineering*, 380–392. doi:10.1016/j.datak.2008.11.002

Hussey, R., Williams, S., & Mitchell, R. (2012). Automatic keyphrase extraction: A comparison of methods. In *Proc. of the International Conference on Information Process, and Knowledge Management* (pp. 18-23).

Ji, L., Sum, M., Lu, Q., Li, W., & Chen, Y. (2007). Chinese terminology extraction using window-based contextual information. In *Proceeding of CICLing* (pp. 62–74). LNCS. doi:10.1007/978-3-540-70939-8_6

Knoth, P., Schmidt, M., Smrz, P., & Zdrahal, Z. (2009). Towards a framework for comparing automatic term recognition methods. In *Proceedings of the Conference Znalosti*.

Krauthammer, M., & Nenadic, G. (2004). Term identification in the biomedical literature. *Journal of Biomedical Informatics*, 512–526. doi:10.1016/j.jbi.2004.08.004 PMID:15542023

Kupsc, A. (2006). Extraction automatique de termes a` partir de textes polonais. *Journal Linguistique de Corpus. LabTestOnLine*. (n.d.). Retrieved from http://labtestsonline.org/

Lossio Ventura, J. A., Jonquet, C., Roche, M., & Teisseire, M. (2013). Combining C-value and keyword extraction methods for biomedical terms extraction. In *Proceedings of the 5th International Symposium on Languages in Biology and Medicine*.

Lv, Y., & Zhai, C. X. (2011). When documents are very long, BM25 fails! In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 1103–1104).

Medelyan, O., Eibe, F., & Witten, I. H. (2009). Human-competitive tagging using automatic keyphrase extraction. In *Proceedings of the International Conference of Empirical Methods in Natural Language Processing (EMNLP),* Singapore.

MeSH (Medical Subject Headings) is the NLM controlled vocabulary thesaurus used for indexing articles for PubMed. (2012). Retrieved from http://www.ncbi.nlm.nih.gov/mesh

Nenadic, G., Spasic, I., & Ananiadou, S. (2003). Morpho-syntactic clues for terminological processing in Serbian. In *Proceedings of the EACL Workshop on Morphological Processing of Slavic Languages* (pp. 79–86).

Noy, N. F., Shah, N. H., Whetzel, P. L., Dai, B., Dorf, M., & Griffith, N. B. et al. (2009). BioPortal: Ontologies and integrated data resources at the click of a mouse. *Nucleic Acids Research*, *37*, 170–173. doi:10.1093/nar/gkp440 PMID:19483092

Robertson, S. E., Walker, S., & Beaulieu, M. (1999). Okapi at TREC-7: Automatic ad hoc, filtering, VLC and interactive track. IN, 21, 253–264.

Sclano, F., & Velardi, P. (2007). TermExtractor: A web application to learn the common terminology of interest groups and research communities. In Enterprise Interoperability II (pp. 287-290).

Spela Vintar. (2004). Comparative evaluation of c-value in the treatment of nested terms. In *Proceedings of the Workshop (Methodologies and Evaluation of Multiword Units in Realworld Applications), LREC* (pp. 54–57).

TreeTagger. (n.d.). Retrieved from www.cis.uni-muenchen.de/~schmid/tools/TreeTagger

Unified Medical Language System (UMLS). (2013). Retrieved from http://www.nlm.nih.gov/research/umls

Zhang, Y., Milios, E., & Zincirheywood, N. (2004). A comparison of keyword and keyterm-based methods for automatic web site summarization. In *Proceedings of the AAAI04 Workshop on Adaptive Text Extraction and Mining* (pp. 15–20).

Zhang, Z., Iria, J., Brewster, C., & Ciravegna, F. (2008). A comparative evaluation of term recognition algorithms. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC08)*.

ENDNOTES

1 http://www.nlm.nih.gov/research/umls
2 http://mesh.inserm.fr
3 http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/
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/