Towards a mixed approach to extract biomedical terms from documents

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

Towards a mixed approach to extract biomedical terms from documents

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

** LIRMM, UM2, CNRS*
Montpellier, FRANCE
e-mail : fName.lName@lirmm.fr

*** Cirad, Irstea, UMR TETIS*
F-34093 Montpellier, FRANCE
e-mail : fName.lName@teledetection.fr

ABSTRACT

The proposed work aims at automatically extracting biomedical terms from free text. We present new extraction methods taking into account linguistic patterns specialized for the biomedical field, 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 we 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.

Keywords: Biomedical Natural Language Processing (BioNLP), Biomedical Term Extraction, Biomedical Thesaurus, Statistic Measure, Text Mining.

INTRODUCTION

Huge amount of biomedical data are now available online, embedding expressions and terms used by the community. These data are often composed of plain text field, e.g., clinical trial description, adverse event report or electronic health records. Although in the biomedical domain hundred of terminologies and ontologies are offered to describe such languages (Musen et al., 2009), they often miss concepts or possible alternative terms for those concepts. Our motivation is thus to improve the precision of automatic term extraction. As language evolves faster than our ability to formalize and catalog it, an automatic and efficient process is needed. This is even truer for French in which the number of formalized terms 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 a field 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*.

	ATR	AKE
	Automatic Term Recognition	Automatic Keyword Extraction
Input	one large corpus	single document of a dataset
Output	technical terms of a domain	keywords that describe the document
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). 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 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 follow. 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. Finally, Section *"Experiments and Results"* details and discusses the conducted experiments and the associated results.

RELATED WORK

ATR proposals 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 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 resources of terminology in order to locate term occurrences in texts. 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 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 for 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 baselines measures as well as new combinations of these measures for automatic biomedical terms extraction. In Subsection A, the defined extensions of the basic measures are detailed. 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 combination of the basic measures: (i) Computing harmonic mean combinations, (ii) Taking into account the Okapi value and TFIDF value within the calculus of C-value.

Our method for automatic term extraction has 4 main steps, described in *Figure 1*: (1) Part of Speech, (2) Candidate terms extraction following patterns, (3) Ranking of candidate terms, (4) Computing of new combination measures. We execute those 4 steps by taking either C-value (right branch) or Okapi/TFIDF (left branch) as baseline method. Notice that as the input of C-value is a unique element and the weighting measure deals with many documents (e.g. Table 1), we need to merge all documents to build a single textual element. An additional first step, not shown in the workflow, is the creation of patterns for both languages.

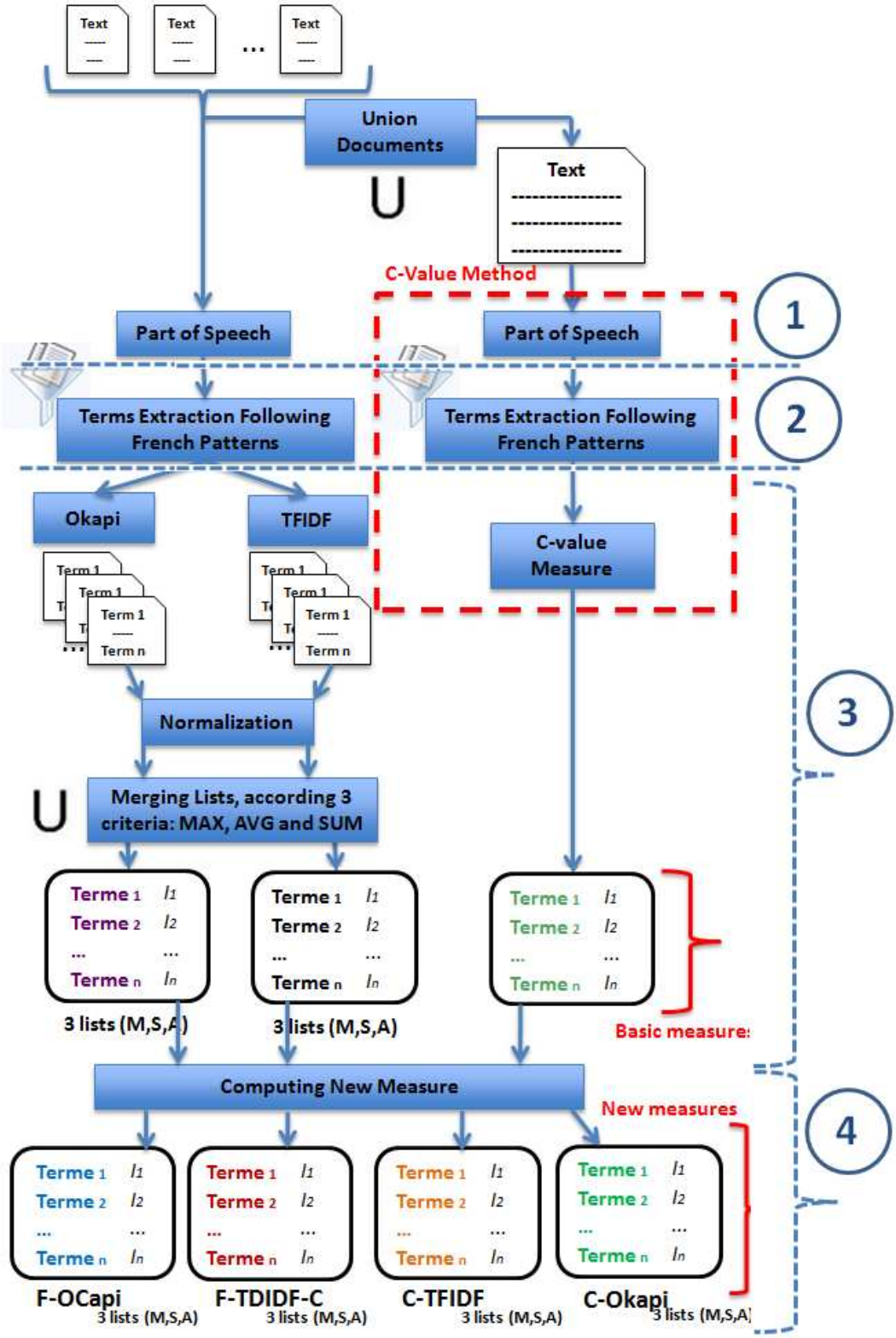

Figure 1: Workflow of Biomedical Term Extraction.

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 (Unified Medical Language System) for English and MeSH (Medical Subject Headings) for French. First, a part-of-speech tagging of the biomedical terms is done by using TreeTagger, a tool for annotating text with part-of-speech and lemma information. 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:*

	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

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

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. Three tools (TreeTagger, Stanford Tagger and Brill's rules) have been compared for this task. TreeTagger is chosen as it gives better results and could be used both for French and English texts.

Term extraction based on biomedical patterns, *see part (2) in Figure 1.*

We only select the terms which linguistic structure is in the pattern lists (English or French). The pattern filtering occurs by language, i.e. when the document is in French, only the 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.

Ranking of candidate terms, *see part (3) in Figure 1.*

A) **Ranking terms with C-value:** The C-value method combines linguistic and statistical information (Frantzi et al., 2000). The linguistic information is associated to general regular expression as 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.

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

Where *a* is the candidate term, |*a*| the number of words in *a*, *f (a)* the frequency of *a* in the document, *Sa* the set of terms that contain *a* and |*Sa*| the number of terms in *Sa*. Basically, C-value either uses frequency of the term (first case of Equation (1)) if the term is not included in other terms. Otherwise, it decreases this frequency if the term appears in other terms, by using the frequency of those other terms (second case of Equation (1)).
We improve the measure in order to extract all terms (single-word + multi-word terms), also proposed in (Barrón et al., 2009) in a different way, in the calculation of the algorithm, ours does not allow null values and is by changing *w(a) = log2(|a| + 1)*. Note that we do not use a stop word list nor a threshold for frequency.
Proposed improvements:
- Linguistic filtering by French/English pattern lists.
- No stop-list.
- No frequency threshold.
- From *w(a) = log2(|a|)* to *w(a)* = log2(|a| + **1**): by adding 1 in the logarithm, that will allow to extract single-word terms as well.

In *Example 1*, the values illustrate the proposed change for the computation of *w(a)* with the original and modified C-value definitions.

	Original C-value	Modified C-value
	w(a) = log2(|a|)	*w(a)* = log2(|a| + **1**)
antiphospholipid antibodies	log2(2)	log2(2 + 1)
white blood	log2(2)	log2(2 + 1)
platelet	**Not possible**	log2(1 + 1)

Example 1: Calculation of w(a).

B) Ranking terms with Okapi - TFIDF: The 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. 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 (Musen

et al., 1999). Okapi can be seen as an improvement of the TFIDF measure, taking into account the document length. Both measures are mostly used for information retrieval and text mining.

- Normalization: The Okapi and TFIDF measures are computed with a variable number of elements, so that the obtained values are not homogeneous. 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.

- Merging lists: It is important 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 it. We merged following three factors: 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 thus obtain 3 lists from Okapi and 3 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.

A) 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.

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

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

B) 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.

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

Where $m_x(a)$ is a weighting measure = $\{Okapi_x, TFIDF_x\}$, and X the factor $\in \{M, S, A\}$. Example 2 shows different ranking of terms with our system based on different measures. This example highlights specific and very relevant terms such as "antiphospholipid antibodies" and "platelet". Indeed these ones obtain a better ranking by using our measures such as F-TFIDF-CM.

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

	Ranking of the terms						
	C-value	***TFIDF***	*Okapi*	***F-TFIDF-C***	*F-OCapi*	*C-TFIDF*	*C-Okapi*
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

Example 2: Rank of terms based on different measures.

DATA AND EXPERIMENTAL PROTOCOL

Test Collection

We used biological laboratory tests as corpus, obtained from Lab Tests Online. This site provides information to patient or family caregiver on clinical lab tests. Each test includes the formal lab test name, its synonyms and many alternate names. Our extracted corpus contains 235 clinical tests (about 400000 words) for English and 137 (about 210000 words) for French.

Validation Data

It is the list of true terms that will be used for the automatic validation. We take the official name, the synonyms and alternate test names of the tests more the UMLS terms for English and the MeSH for French. It allows the evaluation of precision with a proper reference for true terms. Note that as a consequence, the recall is equal to 100% with the whole list of extracted terms.

EXPERIMENTS AND RESULTS

The first evaluation is done with Validation Data, without an expert validation. In order to evaluate automatically that terms present at the top of ranking lists are relevant, we check if they are in biomedical dictionaries (i.e. MeSH and UMLS). The results are given in terms of Precision. Okapi and TFIDF provided three lists (M,S,A). For each combined measure using Okapi or TFIDF, the experiments are conducted with the three lists. So, the number of experiments is equal to 19: C-value(1) + Okapi(3) + TFIDF(3) + F-OCapi(3) + F-TFIDF-C(3) + C-Okapi(3) + C-TFIDF (3). Then, we select all terms (single and multi) or only muli-terms (19 × 2 = 38 experiments for each language). The following sections report part of the experiment results with all and multi terms. In some cases with only the first extracted terms (60, 300 and 900), as it is easier for experts to evaluate only the top-k extracted terms. We evaluated first the basic measures and second with the new combinations 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 3* and *Table 4* show the results of term extraction with $Okapi_X$. Best results were often obtained with $Okapi_M$ for both languages.

	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

Table 3: Precision of $Okapi_X$ on English corpus.

	All Terms			Multi Terms		
	60 terms	300 terms	900 terms	60 terms	300 terms	900 terms
$Okapi_M$	**0,90**	**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 4: Precision of $Okapi_X$ on French corpus.

Table 5 and *Table 6* show the results of terminology extraction with $TFIDF_X$. Best 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.

	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 5: Precision of $TFIDF_X$ on English corpus.

	All Terms			Multi Terms		
	60 terms	300 terms	900 terms	60 terms	300 terms	900 terms
$TFIDF_M$	**0,75**	**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

Table 6: Precision of $TFIDF_X$ on French corpus.

Experiments with C-Value and AKE methods

In this subsection, we evaluated the ATR method, C-value, with the best performances got with AKE methods, i.e. $Okapi_M$ for All and Multi terms, and $TFIDF_M$ for All terms and $TFIDF_S$ for Multi terms. *Table7* and *Table 8* present the results of terminology extraction comparing the best results of basis measures: C-Value, $Okapi_M$, and $TFIDF_{MS}$. The best precision rate is generally obtained with $TFIDF_{MS}$ for English and $Okapi_M$ for French.

	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**
$Okapi_M$	**0,96**	**0,95**	0,82	0,68	0,62	0,55
$TFIDF_M, TFIDF_S$	0,97	0,96	0,84	**0,82**	**0,71**	0,61

Table 7: Precision of C-value, Okapi and TFIDF on English 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**
$Okapi_M$	**0,90**	**0,61**	0,37	**0,53**	0,31	0,37
$TFIDF_M, TFIDF_S$	0,75	0,51	0,37	0,53	0,33	0,22

Table 8: Precision of C-value, Okapi and TFIDF on French corpus.

Experiments with new combined measures

The new measures are evaluated. The first one is based on harmonic mean between the ATR method and AKE methods, $F - OCapi_X$ and $F - TFIDF - C_X$ (see Subsection « A » of *Computing the New Combined Measures*), and the second one, the frequency is replaced by the values obtained from AKE methods, $C - Okapi_X$ and $C - TFIDF_X$ (see Subsection « B » of *Computing the New Combined Measures*). *Table 9* and *Table 10* 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.

	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

Table 9: 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	**0,62**	**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 10: Precision comparison of new measures for French.

Manual validation

In order to have an expert validation, we gave 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 11* and *Table 12* 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 13* and *Table 14* 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.

	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 11: Precision of the best measures for the extraction of all terms for English.

	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 12: Precision of the best measures for the extraction of all terms for French.

	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 13: Precision of $F - TFIDF - C_M$ for English 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%**

Table 14: Precision of $F - OCapi_M$ for French with automatic and manual validations.

DISCUSSION

In the results of AKE methods, TFIDF obtains better results than Okapi. The main reason is associated to the size of the corpus (that is larger) and Okapi works better when the corpus size is not too big (Lv et al., 2011).

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 and F-OCapi.

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 instance, they can represent new terms to add in biomedical dictionaries. So in Tables 13 and 14, the precision rate is naturally higher with a manual validation.

CONCLUSION AND FUTURE WORK

The proposed work defines a new process to automatically extract biomedical terminology for proposing relevant terms to experts. For term ranking, 19 measures have been proposed for two languages, French and English. Conducted experiments have shown that C-value can be used to extract French biomedical terms, which was not stated in the literature before. The precision of the C-value in previous works were only 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. Secondly, 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 conception of a web application & web service that someone can query to use any of our proposed biomedical term extractions methods on other datasets.

REFERENCES

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

Rudi L. Cilibrasi, Paul M. B. Vitanyi. (2007). The Google Similarity Distance. *IEEE Trans. on Knowl. and Data Eng.* pp.370–383.

Nees Jan 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.

Jorge Gracia , Raquel Trillo , Mauricio Espinoza , Eduardo Mena. (2006). Querying the web: a multiontology disambiguation method. *Proceedings of the 6th international conference on Web engineering,* pp.241– 248.

Jorge Gracia , Raquel Trillo , Mauricio Espinoza , Eduardo Mena. (2008). Web-Based Measure of Semant ic Relatedness. *Proceedings of the 9th international conference on Web Information Systems Engineering,* pp.136–150.

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. *Proc. 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. *Proc. of Computer Science and Information Technology (IMCSIT)* pp 411-418.

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

Anna Kupsc. (2006). Extraction automatique de termes a` partir de textes polonais. *Journal Linguistique de Corpus.*

LabTestOnLine, http://labtestsonline.org/

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.

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

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

Goran Nenadic , Irena Spasic , Sophia Ananiadou. (2003). Morpho-syntactic clues for terminological processing in Serbian. *Proc. 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.

S E Robertson, S Walker and M 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.

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

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

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

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 LREC08*.