Biomedical Terminology Extraction: A new combination of Statistical and Web Mining Approaches

Juan Antonio Lossio Ventura[1], Clement Jonquet[1],
Mathieu Roche[1,2], Maguelonne Teisseire[1,2]

[1] LIRMM, Université de Montpellier 2, CNRS – fName.lName@lirmm.fr
[2] Cirad, Irstea, UMR TETIS – fName.lName@teledetection.fr

Abstract

The objective of this work is to combine statistical and web mining methods for the automatic extraction, and ranking of 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 and we investigate which combinations are more relevant for different cases. Each measure gives us a ranked list of candidate terms that we finally re-rank with a new web-based measure. Our experiments show, first 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-word terms; second, that best precision results are often obtained when we re-rank using the web-based measure. We illustrate our results on the extraction of English and French biomedical terms from a corpus of laboratory tests available online in both languages. The results are validated by only using UMLS (in English) and MeSH (in French) as reference dictionary.

Résumé

L'objectif de ce travail est de combiner les méthodes d'extraction statistiques et la fouille du web pour l'extraction automatique et le classement des termes biomédicaux à partir de documents textuels. Nous présentons de nouvelles méthodes d'extraction qui utilisent des patrons linguistiques spécialisés du domaine biomédical associés à des mesures d'extraction de termes, tels que C-value, et des mesures d'extraction de mots-clés comme Okapi BM25 et TFIDF. Nous proposons plusieurs combinaisons de ces mesures afin d'améliorer le processus d'extraction et de classement. Chaque mesure nous donne une liste ordonnée des termes candidats que nous avons finalement réordonnée avec une nouvelle mesure web. Nos expérimentations montrent, d'abord, que la moyenne harmonique de C-value avec une mesure d'extraction de mots-clés offre de meilleurs résultats de précision que leur utilisation seule pour l'extraction des termes composés d'un seul mot et des syntagmes. Les meilleurs résultats de précision sont souvent obtenus quand nous appliquons la mesure fondée sur le web. Nous illustrons nos résultats à partir de l'extraction des termes biomédicaux en anglais et en français sur un corpus de tests de laboratoire disponibles en ligne dans les deux langues. Les résultats sont validés à l'aide de UMLS (en anglais) et MeSH (en français) comme dictionnaire de référence.

Keywords: Biomedical Natural Language Processing (BioNLP), Biomedical Thesaurus, Statistic Measure, Text Mining, Web Mining, C-value.

1. Introduction

The huge amount of data available online today is often composed of plain text fields, for instance clinical trial description, 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 hundreds 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 term extraction processes, 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 formalized in terminologies is significantly less sizeable 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). Similarly, Automatic Keyword Extraction (AKE) is the process of extracting the most relevant words or phrases in a document with the purpose 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**
	Automatic **T**erm **R**ecognition	**A**utomatic **K**eyword **E**xtraction
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*

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 terms and it has never been applied to French text, which is appealing in our case.

Our work follows two main steps: (a) we create new extraction methods by combining in different manners ATR and AKE measures, and we select the best list of ranked candidate terms, (b) we re-rank these extracted lists with a new web-based measure to obtain a new ranked list of candidate terms that maximize precision. Our experiments results present a great improvement of precision with these new combined methods. We give priority to precision in order to focus on the 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 and the web-based measure for re-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. XTRACT (Smadja, 1993) is a statistical method, first it extracts the binary terms located in a window of ten words. The binary terms selected are those that exceed a statistically significantly frequency due to chance. The next step is to extract the terms containing the binary terms found in the previous step. Another method is ACABIT (Daille et al., 1994) that performs a linguistic analysis to convert the nominal terms in binary terms. Then these terms are sorted according to statistical measures. *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, Slovenian, Polish, Chinese (Ji et al., 2007), Spanish (Barrón et al., 2009), and Arabic, 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 them 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 keyword 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 baseline measures and their customizations as well as the new combinations of these measures and the new web-based measure that we propose for automatic biomedical terms extraction and ranking. Our method for automatic term extraction has five main steps; described in figure 1:

(1) Part-of-Speech tagging,
(2) Candidate term extraction,
(3) Ranking of candidate terms,
(4) Computing the new combined measures,
(5) Re-ranking using web-based measure.

We execute those five steps taking either *C-value* (right branch), and *Okapi/TFIDF* (left branch) as the baseline method. Notice that because *C-value* is a method that deals with a single corpus as input whereas the weighting measure deals with several documents (cf. table 1) then we need to do the union of documents of the corpus in the right branch case, in order to consider the whole corpus as a single document. A preliminary step not represented in figure 1 is the design of patterns for French and English, as described hereafter.

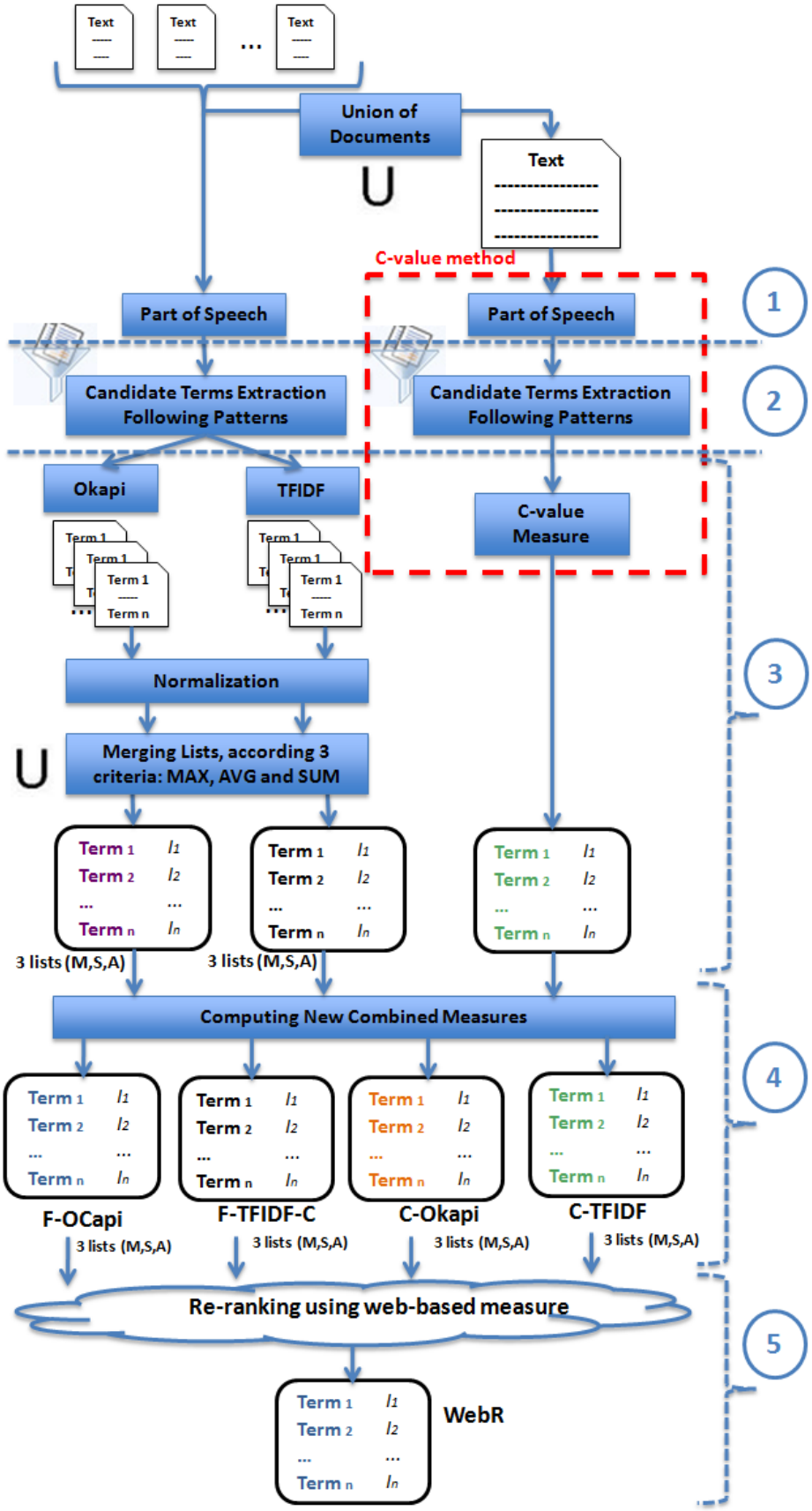

Figure 1. Workflow Methodology for Biomedical Term Extraction.

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 chose TreeTagger which gave the best results and is usable both for French and English.

3.2. Candidate Terms Extraction

3.2.1. Building biomedical patterns

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 to 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. Examples of patterns 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. 9 most frequent syntactic structures of biomedical terms.

3.2.2. Candidate terms extraction following patterns

Before applying any measures we filter out the content of our input corpus using patterns previously computed. We select only the terms whose syntactic structure is in the pattern 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 a linguistic pattern, and the statistical information is the value assigned with the C-value measure based on frequency of

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

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

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

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.

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

	Original *C-value*	Modified *C-value*
	w(a) = log2(|*a*|)	*w(a)* = log2(|*a*| + **1**)
antiphospholipid antibodies	$\log_2(2) = 1$	$\log_2(2 + \mathbf{1}) = 1.6$
white blood	$\log_2(2) = 1$	$\log_2(2 + \mathbf{1}) = 1.6$
platelet	log2(1) = **0**	$\log_2(1 + \mathbf{1}) = 1$

Table 3. Calculation of w(a).

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 in relatively to the corpus it is included in (and relatively to the size of the document in the case of *Okapi*). The output is a ranked list of terms for each document, which is often used in information retrieval, to rank documents depending on 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. In order to manipulate these lists, the weights obtained from each document must be normalized. Once values are normalized we have to merge the terms into a single list for the whole corpus to compare the results. Clearly precision will depend on the method used to perform such merging. We merged the following three functions, which calculate respectively the sum(S), max(M) and average(A) of the measured values of the term in the whole 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 aim of improving the precision of term extraction we have conceived two new combined measures schemes, 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.

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

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

3.4.2. 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 the Okapi or TFIDF values of the terms (over the whole corpus).

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

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

Table 4 shows different rankings of terms with our system based on different measures. This example highlights specific and very relevant terms such as "antiphospholipid antibodies" and "platelet" which are explicitly better ranked with *F-TFIDF-CM* (15,45); in comparison for "white blood" we get the opposite effect (796), because this candidate term is not a biomedical term by itself.

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

Table 4. Ranking of terms based on different measures.

3.5. Re-ranking using Web-based Measure

After the extraction of terms we use a web-based measure to re-rank the candidate terms in order to augment the top *k* terms precision.

Different web mining studies focus on semantic similarity, semantic relatedness (Gracia et al., 2008). It means assessing the degree to which some words or concepts are related, considering not only similarity but any possible semantic relationship among them. They are also used for multi-ontology disambiguation (Gracia et al., 2006). Web-based measures use web search engines to compute this similarity. One of the best-known web-based measures is the Normalized Google Distance (Cilibrasi et al., 2007).

Our web-based measure has for objective to tell us if a candidate term is a real biomedical term or not. It is especially appropriate for multi terms, as it computes the dependence

between the words of a term. In our case, we compute a "strict" dependence, which means the proximity of words of terms (i.e., neighboring words) is calculated with a strict restriction. In comparison with other web-based measures (Cilibrasi et al., 2007), *WebR* reduces the number of pages to consider by taking only into account web pages containing all the words of the terms. In addition, our measure can be easily adopted for all types of multi terms.

$$WebR(a) = \frac{num\_doc("a")}{num\_doc(a)} \tag{5}$$

Where *a* is the candidate term, num_doc("*a*") the number of documents returned by the search engine with exact match only with multi term *a* (query with quotation marks "*a*"), num_doc(*a*) the number of documents returned by the search engine including not exact match (query *a* without quotation marks), i.e., the whole documents containing the words of the multi term *a*. For example, the multi term treponema pallidum, will generate 2 queries, the first num_doc("treponema pallidum") which returns with Yahoo 1'100'000 documents, and the second query num_doc(treponema pallidum) which returns 1'300'000 documents, then:

$$WebR(treponema\ pallidum) = \frac{1\,100\,000}{1\,300\,000} = 0.85$$

In our workflow we have tested Yahoo, Bing and Google, but WebR uses Yahoo because the results were the best. *WebR* re-ranks the list of candidate terms returned by the combined measures. This is the final output of our workflow, on which we can evaluate precision taking the top *k* terms (P@k, in the following k = 60,300,900), and compare them to results obtained either directly by baseline measures or by new combined measures without *WebR*.

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

4. Experiments and Results

4.1. Data and Experimental Protocol

We used biological laboratory tests, extracted from Lab Tests Online[4] as a 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 (at a glance, the test sample, the test and common questions), they are documents that contain free text. Our extracted corpus contains 235 clinical tests (about 400 000 words) for English and 137 (about 210 000 words) for French.

In order to automatically validate our candidate terms we use UMLS for English and MeSH for French as dictionaries.

4.2. Experiments and Results

The evaluation was done automatically. Results are evaluated in terms of *precision* obtained over the top *k* terms (P@k, where k=60, 300, 900) at different steps of our workflow presented in the 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

[4] http://labtestsonline.org/

done with the three lists. Therefore, the number of ranked lists to compare is *C-value*(1) + *Okapi*(3) + *TFIDF*(3) + *F-OCapi*(3) + *F-TFIDF-C*(3) + *C-Okapi*(3) + *C-TFIDF*(3) + *WebR*(1) = 20. In addition we experimented the workflow either for all (single and multi) or multi terms which finally gave 40 ranked lists.

The following paragraphs show part of the experiment results done for all (single- and multi-word terms) or multi terms. In the following we narrow down the presented results by keeping for the next workflow step only the best results.

4.2.1. Results obtained with baselines and new combined measures

	English						**French**					
	All Terms			Multi Terms			All Terms			Multi Terms		
	P@60	P@300	P@900	P@60	P@300	P@900	P@60	P@300	P@900	P@60	P@300	P@900
$Okapi_M$	0.96	0.95	0.82	0.68	0.62	0.54	**0.90**	0.61	0.37	0.53	0.31	0.18
$Okapi_S$	0.83	0.89	0.85	0.58	0.57	0.55	0.30	0.31	0.37	0.23	0.30	0.37
$Okapi_A$	0.72	0.31	0.27	0.48	0.39	0.26	0.52	0.31	0.16	0.30	0.17	0.16
$TFIDF_M$	0.97	0.96	0.84	0.71	0.63	0.54	0.75	0.51	0.37	0.45	0.28	0.18
$TFIDF_S$	0.96	0.95	**0.93**	0.82	0.71	0.61	0.68	0.48	0.42	0.53	0.33	0.22
$TFIDF_A$	0.78	0.74	0.63	0.50	0.40	0.37	0.12	0.39	0.29	0.17	0.16	0.11
C-value	0.88	0.92	0.89	0.72	0.71	0.62	0.43	0.42	0.43	0.35	0.34	**0.26**
$\boldsymbol{F-OCapi_M}$	0.73	0.87	0.84	0.79	0.69	0.58	0.73	**0.62**	**0.43**	**0.65**	**0.35**	0.22
$\boldsymbol{F-TFIDF-C_M}$	**0.98**	**0.97**	0.86	**0.98**	**0.73**	**0.65**	0.85	0.57	0.39	0.62	0.31	0.19
$C-Okapi_S$	0.88	0.86	0.80	0.61	0.58	0.53	0.28	0.32	0.34	0.23	0.28	0.20
$C-TFIDF_S$	0.96	0.95	0.86	0.85	0.71	0.61	0.65	0.55	0.38	0.50	0.32	0.19

Table 5. Extract of precision comparison for term extraction for English and French

Table 5 compares 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. This table proves that the combined measures based on the harmonic mean are better than the baselines measures, and especially 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 *C-Okapi*$_X$ and *C-TFIDF*$_X$ are not better than *Okapi*$_X$ or *TFIDF*$_X$ used directly. The main reason for this is because the performance of those new combined measures is absorbed by the effect of extracting also single word terms. Definitively, all the new combined measures are really performing better for multi word terms.

The results of AKE methods for English show that *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).

In addition, table 5 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%.

4.2.2. WebR results

Our web mining approach is applied at the end of the process. With a small number of terms because the number of queries from an application to the search engines is limited, we took

the lists with best results. The objective is to re-rank the 300 terms of each list putting the "true" terms at the top of the list, in this way the precision by intervals is improved. For this we had to choose the list, which got the best precision in the automatic validation.

Due to the restriction on the number of queries to search engines, it is more interesting for us first to evaluate the web measure with the French data. Table 6 shows the results between *F-OCapi$_M$* and the *WebR* with automatic validation. We can see that *WebR* gets better results by intervals, this means true biomedical terms have a better ranking.

	Multi Terms with Automatic Validation					
	P@30	P@60	P@90	P@120	P@180	P@300
$F - OCapi_M$	63.33%	65.00%	53.33%	49.17%	39.44%	**34.67%**
WebR (Yahoo)	**80.00%**	**68.33%**	**61.11%**	**57.50%**	**47.22%**	**34.67%**

Table 6. Precision comparison between F-OCapi$_M$ and WebR with automatic validation.

4.2.3. Discussion

Several terms proposed by our system are considered 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 actually irrelevant. Indeed, elements that are not automatically validated can be considered relevant after manual validation. For instance, they can represent new terms to add in biomedical ontologies or terminologies. Therefore, we proceed to a manual validation of the rest of the terms (i.e., the ones not found in the validation dictionary). For this, we gave a list of extracted terms to a user to validate manually. Table 7 shows the precision evaluated through human review for the best new combined measures for each language and for the web measure only for French. Note that manual validation confirms that our ranking measure has a good behavior because the precision value is better for first terms. Table 7 also shows clearly that *WebR* gets better results by intervals than *F-Ocapi$_M$*.

		Multi Terms with Manual Validation					
		P@30	P@60	P@90	P@120	P@180	P@300
English	$\boldsymbol{F - TFIDF - C_M}$	**100.00%**	**100.00%**	**99.17%**	**98.89%**	**96.67%**	**93.00%**
French	$F - OCapi_M$	**100.00%**	98.33%	95.56%	95.83%	95.00%	**91.67%**
	WebR (Yahoo)	**100.00%**	**98.33%**	**97.78%**	**97.50%**	**95.56%**	**91.67%**

Table 7. Precision of F-TFIDF-C$_M$ for English and F-OCapi$_M$, WebR for French with ***manual validation.***

We also have done experiments with two more corpora: (i) the Drugs data from MedlinePlus in English and, (ii) PubMed citations' titles in English and French, we have verified that the new combined measures are performing better, particularly those based on the harmonic mean, *F-TFIDF-C$_M$* and *F-OCapi$_M$*.

5. Conclusions and Perspectives

This work presents 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 obtained by combining three existing methods and another new web-based measure. The best results were obtained by combining C-value with the best results from AKE methods, i.e., *F-TFIDF-*C_M and *F-OCapi*$_M$.

Finally, *WebR* was applied to re-rank the best list of candidate terms to move "true" biomedical terms towards the top of the list and thus to improve the P@k. The evaluation shows that *WebR* applied after *F-OCapi*$_M$ got the best precision for the extraction of French term.

The fact that we found false positive means that the term is not found in the validation data set. Then, we proposed a manual validation, for which the new precision results are very good and encouraging to use in terminology enrichment scenarios.

For our future evaluations, we will enrich our dictionaries with BioPortal's[5] terms for English and CISMeF's[6] 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.

Our work shows a comparison with some measures used in the literature, one of our objectives is to compare our work with a large number of measures on ATR for all domains and ATR applied to biomedicine in order to position our methodology regarding the others.

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

Al Khatib K. and Badarneh A. (2010). Automatic extraction of Arabic multi-word terms. *Proceeding of Computer Science and Information Technology*. pp 411-418.

Barrón-Cedeño A., Sierra G., Drouin P. and Ananiadou S. (2009). An Improved Automatic Term Recognition Method for Spanish. *Proceeding of Computational Linguistics and Intelligent Text Processing*, pp 125-136.

Cilibrasi R.L. and Vitanyi P. (2007). The Google Similarity Distance. *IEEE Transactions on Knowledge and Data Engineering*, pp.370–383.

Daille B., Gaussier É. and Langé JM. (1994). Towards Automatic Extraction of Monolingual and Bilingual Terminology. *The 15th International Conference on Computational Linguistics (COLING-94), Kyoto, Japan.*

Eck N., Waltman L., Noyons E. and Buter R. (2010). Automatic term identification for bibliometric mapping. *SpringerLink, Scientometrics*, Volume 82, Number 3.

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

Gaizauskas R., Demetriou G. and Humphreys K. (2000). Term Recognition and Classification in Biological Science Journal Articles. *Proceedings of the Computational Terminology for Medical and Biological Applications Workshop*, pp 37–44.

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

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

Gracia J., Trillo R., Espinoza M. and Mena E. (2006). Querying the web: a multiontology disambiguation method. *Proceedings of the 6th international conference on Web engineering*, pp.241–248.

Gracia J., Trillo R., Espinoza M. and Mena E. (2008). Web-Based Measure of Semantic Relatedness. *Proceedings of the 9th international conference on Web Information Systems Engineering*, pp.136–150.

Hliaoutakis A., Zervanou K. and Petrakis E. (2009). The AMTEx approach in the medical document indexing and retrieval application. Data and Knowledge Eng., pp 380-392.

Hussey R., Williams S. and Mitchell R. (2012). Automatic keyphrase extraction: a comparison of methods. *Proceedings of the International Conference on Information Processing and Knowledge Management*, pp. 18-23.

Ji L., Sum M., Lu Q., Li W. and Chen Y. (2007). Chinese Terminology Extraction Using Window-Based Contextual Information. *Proceeding of CICLing*, LNCS, pp.62–74.

Knoth P., Schmidt M., Smrz P. and Zdrahal Z. (2009). Towards a Framework for Comparing Automatic Term Recognition Methods. *Conference Znalosti*.

Krauthammer M. and Nenadic G. (2004). Term identification in the biomedical literature. *Journal of Biomedical Informatics*, pp 512–526.

Lv Y. and Zhai C. (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.

Medelyan O., Frank E. and Witten I. (2009). Human-competitive tagging using automatic keyphrase extraction. *Proceeding of the International Conference of Empirical Methods in Natural Language Processing*, Singapore.

Mima H. and Ananiadou S. (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

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

Noy N., Shah N., Whetzel P., Dai B., Dorf M., Griffith N., Jonquet C., Rubin D., Storey M., Chute C. and Musen M. (2009). BioPortal: ontologies and integrated data resources at the click of a mouse. *Nucleic Acids Research*, pp. 170-173 vol. 37.

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

Sclano F. and 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.

Smadja F. (1992). Xtract: An overview. Computers and the Humanities 26(5-6): 399-413.

Vintar S. (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.

Zhang Y., Milios E. and Zincirheywood N. (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.

Zhang Z., Iria J., Brewster C. and Ciravegna F. (2008). A Comparative Evaluation of Term Recognition Algorithms. *Proceedings of the Sixth International Conference on Language Resources and Evaluation*.