# Integration of Linguistic and Web Information to Improve Biomedical Terminology Extraction

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

▶ **To cite this version:**

Juan Antonio Lossio-Ventura, Clement Jonquet, Mathieu Roche, Maguelonne Teisseire. Integration of Linguistic and Web Information to Improve Biomedical Terminology Extraction. IDEAS: International Database Engineering & Applications Symposium, Jul 2014, Porto, Portugal. pp.265-269, 10.1145/2628194.2628208 . lirmm-01068547v2

**HAL Id: lirmm-01068547**

**https://hal-lirmm.ccsd.cnrs.fr/lirmm-01068547v2**

Submitted on 30 Sep 2014

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# Integration of Linguistic and Web Information to Improve Biomedical Terminology Extraction

Juan Antonio Lossio-Ventura
LIRMM, CNRS &
University Montpellier 2
Montpellier, France
juan.lossio@lirmm.fr

Clement Jonquet
LIRMM, CNRS &
University Montpellier 2
Montpellier, France
jonquet@lirmm.fr

Mathieu Roche, and Maguelonne Teisseire
Cirad, Irstea, TETIS & LIRMM
Montpellier, France
mathieu.roche@cirad.fr
teisseire@teledetection.fr

## ABSTRACT

Comprehensive terminology is essential for a community to describe, exchange, and retrieve data. In multiple domain, the explosion of text data produced has reached a level for which automatic terminology extraction and enrichment is mandatory. Automatic Term Extraction (or Recognition) methods use natural language processing to do so. Methods featuring linguistic and statistical aspects as often proposed in the literature, solve some problems related to term extraction as low frequency, complexity of the multi-word term extraction, human effort to validate candidate terms. In contrast, we present two new measures for extracting and ranking muli-word terms from domain-specific corpora, covering the all mentioned problems. In addition we demonstrate how the use of the Web to evaluate the significance of a multi-word term candidate, helps us to outperform precision results obtain on the biomedical GENIA corpus with previous reported measures such as *C-value*.

## Categories and Subject Descriptors

H.2.8 [**Database Applications**]: Data mining; I.2.7 [**Natural Language Processing**]: Text analysis

## Keywords

BioNLP, Text Mining, Web Mining, Automatic Term Extraction

## 1. INTRODUCTION

The amount of textual documents available on the web is always growing. Nowadays, analysis on such data for finding interesting knowledge is still a challenge. It is even more true when concepts or terms are domain-based (clinical trial description, adverse event report, electronic health records, customer complaint emails or engineers' repair notes [10]) as it needs more complex approaches for efficiency purpose. Recently, in the Natural Language Processing community, the automatic extraction of representative patterns from plain texts has been addressed to identify terms or expressions from a given corpus such as the Automatic Term Extraction (ATE), or Automatic Term Recognition methods. The associated applications are lexicon update, domain ontology construction, summarization. Term extraction methods usually involves two main steps. The first step extracts candidates by unithood calculation to qualify a string as a valid term. The second step verifies them through termhood measures to validate their domain specificity. Formally, unithood refers to the degree of strength or stability of syntagmatic combinations and collocations, and termhood is defined as the degree that a linguistic unit is related to domain-specific concepts [12]. Such terms may be: (i) single-word terms (usually simple to extract), or (ii) multi-word terms (hard to extract). In this paper, we focus on mutli-word term extraction.

There are some known issues with ATE such as: (i) extraction of non valid terms (noise) or missing of relevant terms with low frequency (silence), (ii) extraction of multi-word terms that inevitably have complex and various structures, (iii) human effort spent for validating the candidate terms, which is usually manually performed [5], (iv) application to large-scale corpora. In response to the above problems, we propose two new measures, the first one called *LIDF-value*, which is statistic- and linguistic-based measure, which deals with first, second and last previously mentioned issues. And the second one called *WAHI*, which is a web-based measure which deals with the three firsts issues. The novelty of the *LIDF-value* measure is an enhance consideration of term unithood, by computing a degree of quality for term unithood. The novelty of the *WAHI* measure is to be web-based which has never been applied within ATE approaches. In this paper, we compare the quality of the proposed methods with the most used baseline measures despite the difficulty in comparing the ATE measures due to the size of the used corpora, due to the absence of libraries implementing methodologies proposed before. We demonstrate that the use of these two measures improves the automatic extraction of domain-specific terms from text collections that do not offer reliable frequency statistical evidence.

The rest of the paper is organized as follow: We discuss related work in Section 2. The two new measures are described in Section 3. Evaluation of our approach is presented in Section 4 followed with conclusions in Section 5.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from Permissions@acm.org.
*IDEAS'14* July 07 - 09, 2014, Porto, Portugal
Copyright 2014 ACM 978-1-4503-2627-8/14/07 $15.00.
http://dx.doi.org/10.1145/2628194.2628208

## 2. RELATED WORK

Several recent studies focused on multi-word (n-grams) and single-word (unigrams) term extraction; existing term extraction techniques can be divided in four broad categories: (i) *linguistic*, (ii) *statistical*, (iii) *machine learning*, and (iv) *hybrid*. All of these techniques are enclosed in text mining approaches. Existing web techniques have not been applied to ATE, but as we will see, such techniques can be adapted for such purpose.

## 2.1 Text Mining approaches

**Linguistic approaches:** Linguistic approaches attempt to recover terms thanks to the formation of syntactic patterns. The main idea is to build rules in order to describe naming structures for different classes using orthographic, lexical, or morphosyntactic characteristics, for instance [8]. As explained in [14], the main approach is to (typically manually) develop rules that describe common naming structures for certain term classes using either orthographic or lexical clues, or more complex morpho-syntactic features.

**Statistical methods:** The statistical techniques chiefly rely on external evidence presented through surrounding (contextual) information. Statistical approaches mainly address the recognition of general terms [21]. The most basic measure is frequency. For instances, *term frequency (tf)* counts the frequency of a term in the corpus; *document frequency (df)* counts the number of documents where term occurs; *average term frequency (atf)*, which is $\frac{tf}{df}$. A comparable research topic, called Automatic Keyword Extraction (AKE), identifies the processes of extracting the most relevant words or phrases in a document. AKE measures are often use for automatic indexing. Keywords, which we define as a sequence of one or more words, provide a compact representation of a document content. IN previous publications, we demonstrated that AKE measures can be adapted for term extraction from a corpus and obtained significant results [15] [16]. We have taken two popular AKE measures, *Okapi BM25* and *TF-IDF* (also called weighting measures) to extract biomedical terms. One might also mention *residual inverse document frequency (RIDF)*, which compares *df* to another model of chance where terms with a particular term frequency are distributed randomly throughout the collection. In addition, *Chi-square* [17] computes how selectively words and phrases co-occur within the same sentences as a particular subset of frequent terms in the document text, it is applied to determine the bias of word co-occurrences in the document text which is then used to rank words and phrases as keywords of the document. Finally, *RAKE* [20] hypotheses that keywords usually consist of multiple words and usually do not contain punctuation or stop words and uses the information of word co-occurrences to determine the keywords.determine the keywords.

**Machine Learning:** Machine Learning systems are usually designed for a specific class of entities and integrate term recognition and term classification. Machine Learning systems use training data to learn useful features for term recognition and classification, but the existence of reliable training resources is one of the main problems as they are not widely available. Some proposed ATE approach uses machine learning [6] [23] [18]. Although machine learning may also generate noise and silence, it facilitates the use of a number of TCs and their features, but the main challenge is to select a set of discriminating features that can be used for accurate recognition (and classification) of term instances.

**Hybrid methods:** Several approaches combine different methods (typically linguistic and statistical) for term extraction. *GlossEx* [13] is a method that considers the probability of the word in a domain corpus divided by the probability of the same word in a general corpus, moreover, the importance of the word is increased according to its frequency in the domain corpus. *Weirdness* [1] considers that the distribution of words in a specific domain corpus is different from the word distribution in a general corpus. *C/NC-value* [7], combines statistical and linguistic information for the extraction of multi-word and nested terms, it the most well known in the literature. 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 [9] [15]. In [24], authors show that *C-value* has the best results compared to other measures cited above. Another measure is *F-TFIDF-C* [16] that combines an ATE measure *(C-value)* and an AKE measure *(TF-IDF)* to extract terms obtaining better results than *C-value*. Also, the *C-value* measure was also applied to many different languages besides English, such as Japanese, Serbian, Slovenian, Polish, Chinese [11], Spanish [2], Arabic, and French [15]. That is why we take *C-value* and *F-TFIDF-C* as baselines for comparison in our experiments.

## 2.2 Web Mining approaches

Different web mining studies focus on semantic similarity, semantic relatedness. It means to quantify the degree in which some words are related, considering not only similarity but any possible semantic relationship among them. The word association measures can be divided in three categories [3]: (i) *Co-occurrence measures* that rely on co-occurrence frequencies of both words in a corpus, (ii) *Distributional similarity-based measures* that characterize a word by the distribution of other words around it, and (iii) *Knowledge-based measures* that use knowledge-sources like thesauri, semantic networks, or taxonomies. In this paper, we focus on co-occurrence measures, because our goal is to extract multi-word terms and we suggest to compute a degree of association between words composing a term. Word association measures are used in several domains like ecology, psychology, medicine, and language processing and recently studied in [19] [22], such as *Dice*, *Jaccard*, *Overlap*, *Cosine*. Another measure to compute the association between words using web search engines results is the Normalized Google Distance [4], which relies on the number of times words co-occur in the document indexed by an information retrieval system. In this study, we compare our results obtained with our web-based measure with basic measures (*Dice*, *Jaccard*, *Overlap*, *Cosine*).

# 3. OUR APPROACH

## 3.1 A new ranking measure based on linguistic and statistical information

### *3.1.1 Linguistic-based measure*

The objective of this measure is to give a higher importance to the term unithood than the measures cited in the state-of-the-art, in order to detect terms that have low frequency. 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.

As previously cited work, we suppose terms of a domain have similar syntactic structure. Therefore, we build a list of the most common linguistic patterns according the syntactic structure of technical terms present in a dictionary, in our case, UMLS[1], which is a set of references biomedical terminologies. We do part-of-speech tagging of the domain dictionary using the Stanford CoreNLP API (POS tagging)[2], then compute the frequency of syntactic structures. We then choose the 200 highest frequencies to build the list of patterns and we compute the following weight: the probability a candidate term may have of being a domain term if its syntactic structure appears in the linguistic pattern list . The number of terms used to build these lists of patterns was 2 300 000. Table 1 illustrates the computation of the linguistic pattern probability.

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

[2] http://nlp.stanford.edu/software/corenlp.shtml

| Pattern | Frequency | Probability |
|---|---|---|
| NN IN JJ NN IN JJ NN | 3006 | 3006/4113 = 0.73 |
| NN CD NN NN NN | 1107 | 1107/4113 = 0.27 |
| | 4113 | 1.00 |

**Table 1: Example of pattern construction (where *NN* is a noun, *IN* a preposition or subordinating conjunction, *JJ* an adjective, and *CD* a cardinal number)**

### *3.1.2 Statistical-based measure*

The objective of our measure called *LIDF-value* (***L**inguisitic patterns*, ***IDF***, and *C-**value*** information) is to compute the termhood for each term, using the $probability$ calculated previously, also with the $idf$, and the *C-value* of each term. The inverse document frequency $(idf)$ is a measure of whether the term is common or rare across all documents. It is obtained by dividing the total number of documents by the number of documents containing the term, and then taking the logarithm of that quotient. The $probability$ and the $idf$ improve the extraction of terms with low frequency. In addition, the *C-value* measure is based on term frequency of terms. The aim of the *C-value* is to improve the extraction of nested terms, i.e., this criteria favors a candidate term that not appears often in a longer term. For instance, in a specialized corpus (Ophthalmology), the Ananiadou et al. [7] found the irrelevant term *soft contact* while the frequent and longer term *soft contact lens* is relevant.

We combined these different statistical information (i.e., $probability$ of linguisitic patterns, $C\text{-}value$, $idf$) to propose a global ranking measure called *LIDF-value* (formula 1). $A$ represents multi-word term; $\mathrm{P}(A_{LP})$ is the probability of $A$ which has the same linguistic structure of linguistic pattern $LP$, it means the weight of the linguistic pattern $LP$ computed in Subsection 3.1.1.

$$LIDF\text{-}value(A) = \mathrm{P}(A_{LP}) \times idf(A) \times C\text{-}value(A) \quad (1)$$

So, to compute *LIDF-value* we execute three step:

(1) **Part-of-Speech tagging:** we apply part-of-speech to the whole corpus, after we take the lemma of each word.
(2) **Candidate terms extraction:** before applying any measures we filter out the content of our input corpus using patterns previously computed. We select only the terms which syntactic structure is in the patterns list.
(3) **Ranking of candidate terms:** finally we compute the value of *LIDF-value* for each term.

In order to improve the ranking, we propose, in the following subsection, to take into account web information to highlight relevant terms.

## 3.2 A new Web ranking measure

Previous studies of web mining approaches query the Web via search engines to measure association of words. This enables to measure the association of words composing a term (e.g., 'soft', 'contact', and 'lens' that compose the relevant term 'soft contact lens'). To measure this association, our web-mining approach takes into account the number of pages provided by search engines (i.e., number of hits). Our Web-based measure has for objective to re-rank the list obtained previously with the *LIDF-value* measure. We will show that it enables improving the precision of the $k$ first terms extracted (see Section 4) and that it is specially appropriated for multi-word term extraction. The Dice's coefficient used in our approach computes some kind of relationship between two words called a *co-occurrence*. This measure is defined by the following formula:

$$D(x,y) \quad = \quad \frac{2 \times P(x,y)}{P(x) + P(y)} \quad (2)$$

Formula 2 leads directly to formula 3.[3] The $nb$ function used in formula 3 represents the number of pages provided by search engines (i.e., Yahoo and Bing). With this measure, we compute a *strict* dependence (i.e., neighboring words by using the operator ' " ' of search engines). For instance, $x$ might represent the word 'soft' and $y$ the word 'contact' in order to calculate the association measure of 'soft contact' term.

$$Dice(x,y) \quad = \quad \frac{2 \times nb(\text{“}x\ y\text{”})}{nb(x) + nb(y)} \quad (3)$$

In a natural way, we extend this approach to $n$ elements as follows:

$$Dice(a_1, ..., a_n) \quad = \quad \frac{n \times nb(\text{“}a_1\ ...\ a_n\text{”})}{nb(a_1) + ... + nb(a_n)} \quad (4)$$

This measure enables to calculate a score for all multi-word terms, such as 'soft contact lens'.

Moreover, we associate Dice criteria with another association measure called $WebR$ (see formula 5 and [16]). This one takes only into account the number of web pages containing all the words of the terms by using operators *" "* and *AND*.

$$WebR(A) = \frac{nb(\text{“}A\text{”})}{nb(A)} \quad (5)$$

Where $A$ = multi-word term, $a_i \in A$ and $a_i = \{noun, adjective, foreign\ word\}$.

For the example of term 'soft contact lens', the numerator corresponds to the number of web pages with the query *"soft contact lens"*, and for the denominator we consider the query *soft AND contact AND lens*.

Finally the global ranking approach combining Dice and $WebR$ is given by $WAHI$ measure (**W**eb **A**ssociation based on **H**its **I**nformation):

$$WAHI(A) = \frac{n \times nb(\text{“}A\text{”})}{\sum_{i=1}^{n} nb(a_i)} \times \frac{nb(\text{“}A\text{”})}{nb(A)} \quad (6)$$

We will show that open-domain (general) resources, such as web, can be exploited to support domain-specific term extraction. Thus, they can be used to compensate for the unavailability of domain-specific resources.

# 4. EXPERIMENTS AND RESULTS

## 4.1 Data, Protocol, and Validation

We use GENIA[4] corpus, which is made up of the 2 000 titles and abstracts of journal articles that have been taken from the Medline database, with more than 400 000 words. GENIA corpus contains linguistic expressions referring to entities of interest in molecular biology such as proteins, genes and cells. The GENIA technical term annotation covers the identification of physical biological entities as well as other important terms. Whereas the Medline indexes broad range of academic articles covering the general or specific domains of life sciences, GENIA is intended to cover a smaller

[3] by writing $P(x) = \frac{nb(x)}{nb_total}$, $P(y) = \frac{nb(y)}{nb_total}$, $P(x,y) = \frac{nb(x,y)}{nb_total}$

[4] `http://www.nactem.ac.uk/genia/genia-corpus/term-corpus`

subject domain: biological reactions concerning transcription factors in human blood cells. Then in order to automatically validate and to cover medical terms, we create a dictionary that contains every term of UMLS as well as all the GENIA technical term. We can now evaluate precision with a proper reference for valid terms.

## 4.2 Results

Results are evaluated in terms of *precision* obtained over the top $k$ terms ($P@k$) for the two measures presented in previous section. The following sections show part of the experiment results done for multi-word terms and considering the top $k$ extracted terms. In the following, we keep only the first 1000 extracted terms.

### 4.2.1 LIDF-value results

Table 2 compares the results of *C-value*, *F-TFIDF-C*, with our new measure *LIDF-value*. Best precision results were obtained with *LIDF-value* whatever $k$. This table proves that *LIDF-value* is better than the baseline measures with a gain in precision that reaches 11% for the first hundred extracted multi-word terms. This result is particularly positive because it is often more interesting to extract multi-word terms because is quite complex.

| | *C-value* | *F-TFIDF-C* | ***LIDF-value*** |
|---|---|---|---|
| P@100 | 0.730 | 0.770 | **0.840** |
| P@200 | 0.715 | 0.725 | **0.790** |
| P@300 | 0.730 | 0.723 | **0.780** |
| P@400 | 0.697 | 0.705 | **0.757** |
| P@500 | 0.674 | 0.694 | **0.752** |
| P@600 | 0.670 | 0.687 | **0.765** |
| P@700 | 0.661 | 0.686 | **0.761** |
| P@800 | 0.644 | 0.669 | **0.755** |
| P@900 | 0.641 | 0.649 | **0.757** |
| P@1000 | 0.635 | 0.637 | **0.746** |
| P@2000 | 0.601 | 0.582 | **0.708** |
| P@5000 | 0.530 | 0.513 | **0.625** |
| P@10000 | 0.459 | 0.439 | **0.574** |
| P@20000 | 0.382 | 0.335 | **0.416** |

**Table 2: Precision comparison with baseline measures**

### 4.2.2 Web Mining results

Our web mining approach is applied at the end of the process, with only the first 1 000 terms extracted during the previous linguistic and statistic measures. The main reason for this limitation is the limited number of automatic queries one can make to search engines. At this step, the objective is to re-rank the 1 000 terms trying to improve the precision by intervals. Each measure listed in Table 3 and Table 4 shows the precision obtained after re-ranking. We experimented *WAHI* with *Yahoo* and *Bing* search engines. Table 3 and Table 4 prove that *WAHI* (either using *Yahoo* or *Bing*) is well adapted for ATE and this measure obtains better precision results than the baselines measures for word association.

### 4.2.3 Summary

*LIDF-value* obtains the best precision results for multi-word term extraction and for each index term extraction ($n$-gram) and for intervals. Table 5 presents the precision comparison of our two measures. In terms of overall precision, our results produce consistent results from GENIA corpus. *WAHI* based on Yahoo obtains best precision (i.e., 90%) for the first $P@100$. In comparison *WAHI* based on Bing obtains a precision of 80%. For the other interval, Table 5 shows that in general *WAHI* based on Bing has the best results. That is very encouraging, because it helps to alleviate the manual validation of candidate terms. The performance of *WAHI* depends of search engine because the associated are different. Then the number of hits returned by these will be different for each case. Moreover, Table 5 highlights that re-ranking with *WAHI* enables to increase the precision of *LIDF-value*. The purpose for which this web-mining measure was created, has been reached.

| | *WAHI* | *Dice* | *Jaccard* | *Cosine* | *Overlap* |
|---|---|---|---|---|---|
| P@100 | **0.900** | 0.720 | 0.720 | 0.76 | 0.730 |
| P@200 | **0.800** | 0.775 | 0.770 | 0.740 | 0.765 |
| P@300 | **0.800** | 0.783 | 0.780 | 0.767 | 0.753 |
| P@400 | **0.800** | 0.770 | 0.765 | 0.770 | 0.740 |
| P@500 | **0.820** | 0.764 | 0.754 | 0.762 | 0.738 |
| P@600 | **0.767** | 0.748 | 0.740 | 0.765 | 0.748 |
| P@700 | **0.786** | 0.747 | 0.744 | 0.747 | 0.757 |
| P@800 | **0.775** | 0.752 | 0.7463 | 0.740 | 0.760 |
| P@900 | **0.756** | 0.749 | 0.747 | 0.749 | 0.747 |
| P@1000 | **0.746** | **0.746** | **0.746** | **0.746** | **0.746** |

**Table 3: Precision comparison of *WAHI with YAHOO* and word association measures**

| | *WAHI* | *Dice* | *Jaccard* | *Cosine* | *Overlap* |
|---|---|---|---|---|---|
| P@100 | **0.800** | 0.740 | 0.730 | 0.680 | 0.650 |
| P@200 | **0.800** | 0.775 | 0.775 | 0.735 | 0.705 |
| P@300 | **0.800** | 0.770 | 0.763 | 0.740 | 0.713 |
| P@400 | **0.800** | 0.765 | 0.765 | 0.752 | 0.712 |
| P@500 | **0.800** | 0.760 | 0.762 | 0.758 | 0.726 |
| P@600 | **0.817** | 0.753 | 0.752 | 0.753 | 0.743 |
| P@700 | **0.814** | 0.7514 | 0.751 | 0.733 | 0.749 |
| P@800 | **0.775** | 0.745 | 0.747 | 0.741 | 0.754 |
| P@900 | **0.778** | 0.747 | 0.748 | 0.742 | 0.748 |
| P@1000 | **0.746** | **0.746** | **0.746** | **0.746** | **0.746** |

**Table 4: Precision comparison of *WAHI with BING* and word association measures**

| | *LIDF-value* | *WAHI* (*Bing*) | *WAHI* (*Yahoo*) |
|---|---|---|---|
| P@100 | 0.840 | 0.800 | **0.900** |
| P@200 | 0.790 | **0.800** | **0.800** |
| P@300 | 0.780 | **0.800** | **0.800** |
| P@400 | 0.757 | **0.800** | **0.800** |
| P@500 | 0.752 | 0.800 | **0.820** |
| P@600 | 0.765 | **0.817** | 0.767 |
| P@700 | 0.761 | **0.814** | 0.786 |
| P@800 | 0.755 | **0.775** | **0.775** |
| P@900 | 0.757 | **0.778** | 0.756 |
| P@1000 | **0.746** | **0.746** | **0.746** |

**Table 5: Precision comparison *LIDF-value* and *WAHI***

# 5. CONCLUSIONS AND FUTURE WORK

The paper presents two measures for automatic multi-word term extraction. The first one, a linguistic- and statistic- measure, *LIDF-value*, improves precision of automatic term extraction in comparison with the most popular term extraction measure. Our approach overcomes the lack of frequency information with the values of *linguistic pattern probability* and *idf*. The second one is a web-based measure. This measure called *WAHI* takes as input the list of terms obtained with *LIDF-value*. *WAHI* enables to reduce the huge human effort for validating candidate terms. Our experiments re-

veal that *LIDF-value* outperforms a state-of-the-art reference measures for extracting terms in the biomedical domain, at least on the GENIA corpus. We experimentally show *LIDF-value* returns best results in comparison with baseline measures (i.e. *C-value* and *F-TFIDF-C*) performing as well for index term extraction ($n$-gram). Moreover our experimental evaluations reveal that *WAHI* improves the results given with *LIDF-value*. A future extension of this work consists in using the Web to extract more terms than those extracted. Moreover, we project to test this general approach on other domains, such as ecology and agronomy. Finally we plan to experiment our proposals on French and Spanish corpora as well as on other English corpora.

## 6. 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, CNRS and IBC project.

## 7. REFERENCES

[1] K. Ahmad, L. Gillam, and L. Tostevin. University of surrey participation in trec8: Weirdness indexing for logical document extrapolation and retrieval (wilder). In *TREC*, 1999.

[2] A. Barron-Cedeno, G. Sierra, P. Drouin, and S. Ananiadou. An improved automatic term recognition method for spanish. In *Computational Linguistics and Intelligent Text Processing*, pages 125–136. Springer, 2009.

[3] D. L. Chaudhari, O. P. Damani, and S. Laxman. Lexical co-occurrence, statistical significance, and word association. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 1058–1068, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.

[4] R. L. Cilibrasi and P. M. Vitanyi. The google similarity distance. *Knowledge and Data Engineering, IEEE Transactions on*, 19(3):370–383, 2007.

[5] M. S. Conrado, T. A. Pardo, and S. O. Rezende. Exploration of a rich feature set for automatic term extraction. In *Advances in Artificial Intelligence and Its Applications*, volume 8265 of *Lecture Notes in Computer Science*, pages 342–354. Springer Berlin Heidelberg, 2013.

[6] J. Foo and M. Merkel. Using machine learning to perform automatic term recognition. In *Proceedings of the LREC 2010 Workshop on Methods for automatic acquisition of Language Resources and their evaluation methods*, pages 49–54, 2010.

[7] K. Frantzi, S. Ananiadou, and H. Mima. Automatic recognition of multi-word terms: the c-value/nc-value method. *International Journal on Digital Libraries*, 3(2):115–130, 2000.

[8] R. Gaizauskas, G. Demetriou, and K. Humphreys. Term recognition and classification in biological science journal articles. In *Proceeding of the Computional Terminology for Medical and Biological Applications Workshop of the 2 nd International Conference on NLP*, pages 37–44, 2000.

[9] A. Hliaoutakis, K. Zervanou, and E. G. Petrakis. The amtex approach in the medical document indexing and retrieval application. *Data & Knowledge Engineering*, 68(3):380–392, 2009.

[10] A. Ittoo and G. Bouma. Term extraction from sparse, ungrammatical domain-specific documents. *Expert Systems with Applications.*, 40(7):2530–2540, June 2013.

[11] L. Ji, M. Sum, Q. Lu, W. Li, and Y. Chen. Chinese terminology extraction using window-based contextual information. In *Proceedings of the 8th International Conference on Computational Linguistics and Intelligent Text Processing (CICLing07)*, pages 62–74, Berlin, Heidelberg, 2007. Springer-Verlag.

[12] K. Kageura and B. Umino. Methods of automatic term recognition: A review. *Terminology*, 3(2):259–289, 1996.

[13] L. Kozakov, Y. Park, T. Fin, Y. Drissi, N. Doganata, and T. Confino. Glossary extraction and knowledge in large organisations via semantic web technologies. In *Proceedings of the 6th International Semantic Web Conference and the 2nd Asian Semantic Web Conference*, 2004.

[14] M. Krauthammer and G. Nenadic. Term identification in the biomedical literature. *Journal of Biomedical Informatics*, 37(6):512–526, December 2004.

[15] J. A. Lossio-Ventura, C. Jonquet, M. Roche, and M. Teisseire. Combining c-value and keyword extraction methods for biomedical terms extraction. In *Proceedings of the Fifth International Symposium on Languages in Biology and Medicine (LBM13)*, pages 45–49, Tokyo, Japan, December 2013.

[16] J. A. Lossio-Ventura, C. Jonquet, M. Roche, and M. Teisseire. Biomedical terminology extraction: A new combination of statistical and web mining approaches. In *Proceedings of Journées internationales d'Analyse statistique des Données Textuelles (JADT2014)*, Paris, France, June 2014.

[17] Y. Matsuo and M. Ishizuka. Keyword extraction from a single document using word co-occurrence statistical information. *International Journal on Artificial Intelligence Tools*, 13(01):157–169, 2004.

[18] D. Newman, N. Koilada, J. H. Lau, and T. Baldwin. Bayesian text segmentation for index term identification and keyphrase extraction. In *Proceedings of 24th International Conference on Computational Linguistics (COLING)*, pages 2077–2092, Mumbai, India, December 2012.

[19] P. Pantel, E. Crestan, A. Borkovsky, A.-M. Popescu, and V. Vyas. Web-scale distributional similarity and entity set expansion. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '09, pages 938–947, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics.

[20] S. Rose, D. Engel, N. Cramer, and W. Cowley. Automatic keyword extraction from individual documents. *Text Mining: Theory and Applications*, pages 1–20, 2010.

[21] N. J. Van Eck, L. Waltman, E. C. Noyons, and R. K. Buter. Automatic term identification for bibliometric mapping. *Scientometrics*, 82(3):581–596, 2010.

[22] R. B. Zadeh and A. Goel. Dimension independent similarity computation. *Journal of Machine Learning Research*, 14(1):1605–1626, January 2013.

[23] X. Zhang, Y. Song, and A. Fang. Term recognition using conditional random fields. In *International Conference on Natural Language Processing and Knowledge Engineering (NLP-KE)*, pages 1–6. IEEE, 2010.

[24] Z. Zhang, J. Iria, C. Brewster, and F. Ciravegna. A Comparative Evaluation of Term Recognition Algorithms. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC08)*, Marrakech, Morocco, May 2008.