# Yet Another Ranking Function for Automatic Multiword Term Extraction

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

[1] University of Montpellier 2, LIRMM, CNRS - Montpellier, France
{juan.lossio,clement.jonquet}@lirmm.fr
[2] Irstea, CIRAD, TETIS - Montpellier, France
mathieu.roche@cirad.fr, maguelonne.teisseire@teledetection.fr

**Abstract.** Term extraction is an essential task in domain knowledge acquisition. We propose two new measures to extract multiword terms from a domain-specific text. The first measure is both linguistic and statistical based. The second measure is graph-based, allowing assessment of the importance of a multiword term of a domain. Existing measures often solve some problems related (but not completely) to term extraction, e.g., noise, silence, low frequency, large-corpora, complexity of the multiword term extraction process. Instead, we focus on managing the entire set of problems, e.g., detecting rare terms and overcoming the low frequency issue. We show that the two proposed measures outperform precision results previously reported for automatic multiword extraction by comparing them with the state-of-the-art reference measures.

## 1 Introduction

The huge amount of data available online today is often composed of plain text fields, e.g., clinical trial descriptions, adverse event reports, electronic health records [14], customer complaint emails or engineers' repair notes [9]. These texts are often written with a specific language (expressions and terms) used by the associated community. There is thus a need for formalization and cataloguing of these technical terms or concepts. But this task is very time consuming.

Automatic Term Extraction (ATE) or Automatic Term Recognition aim to automatically extract technical terminology from a given corpus. Technical terminology is a set of terms used in a domain. Therefore term extraction is an essential task in domain knowledge acquisition, because the technical terminology can be used for lexicon update, domain ontology construction, summarization, named entity recognition, information retrieval. Technical terms are useful to gain further insight into the conceptual structure of a domain. These may be: (i) single-word terms (simple), or (ii) multiword terms (complex). The proposed work focuses on mutliword term extraction.

Term extraction methods usually involve two main steps. The first step extracts candidates by unithood calculation to qualify a string as a valid term.

A. Przepiórkowski and M. Ogrodniczuk (Eds.): PolTAL 2014, LNAI 8686, pp. 52–64, 2014.

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 to which a linguistic unit is related to domain-specific concepts [11]. ATR has been applied to several domains, e.g., biomedical [13] [14] [6] [25] [17], ecological [4], mathematical [22], social networks [15], banking [5], natural sciences [5], information technology [17], and legal.

There are some well-known ATE issues such as: (i) extraction of non-valid terms (noise) or omission of terms with low frequency (silence), (ii) extraction of multiword terms having complex and various structures, (iii) manual validation efforts of the candidate terms [4], and (iv) management of large-scale corpora.

In response to the above problems, two new measures are proposed in this paper. The first one, called *LIDF-value*, is a statistical- and linguistic-based measure and addresses issues i), ii) and iv). The second one, called *TeRGraph*, is a graph-based measure and deals with issues i), ii) and iii). The main contributions are: (1) enhanced consideration of the term unithood, by computing a degree of quality for the term unithood, and, (2) the consideration of the term dependence in the ATE process. The quality of the proposed method is underlined by comparing the results obtained with the most commonly used baseline measures. The experiments were conducted despite difficulties in comparing ATE measures, mainly because of the size of the corpora used, and the lack of available libraries associated with previous works. Our two measures improve the process of automatic extraction of domain-specific terms from text collections that do not offer reliable statistical evidence.

The paper is organized as follows. We first discuss related work in Sect. 2. Then, the two new term extraction measures are detailed in Sect. 3. Precision evaluation is presented in Sect. 4 followed by the conclusions in Sect. 5.

## 2 Related Work

Recent studies have focused on multiword (n-grams) and single-word (unigrams) term extraction. Term extraction techniques can be divided into four broad categories: (i) *Linguistic*, (ii) *Statistical*, (iii) *Machine Learning*, and (iv) *Hybrid*. All of these techniques are encompassed in Text Mining approaches. Graph-based approaches have not yet been applied to ATE, although they have been successively adopted in other Information Retrieval fields and they could be suitable for our purpose.

### 2.1 Text Mining Approaches

**Linguistic Approaches.** These techniques attempt to recover terms via pattern formation. This involves building rules to describe naming structures for different classes by using orthographic, lexical, or morphosyntactic characteristics, e.g., [7]. The main approach is to (typically manually) develop rules describing common naming structures for certain term classes using orthographic or lexical clues, or more complex morpho-syntactic features.

**Statistical Methods.** Statistical techniques chiefly rely on external evidence presented through surrounding (contextual) information. Such approaches are mainly focused on the recognition of general terms [23]. The most basic measures are based on frequency. For instance: *term frequency (tf)* counts the frequency of a term in the corpus; *document frequency (df)* counts the number of documents where a term occurs. A similar research topic, called Automatic Keyword Extraction (AKE), proposes to extract the most relevant words or phrases in a document using automatic indexation. Keywords, which we define as a sequence of one or more words, provide a compact representation of the document's content. Such measures can be adapted to extract terms from a corpus as well as ATE measures. In [14] [13], two popular AKE measures, *Okapi BM25* and *TF-IDF* (also called weighting measures), are used to automatically extract biomedical terms; *residual inverse document frequency (RIDF)* compares the document frequency to another chance model where terms with a particular term frequency are distributed randomly throughout the collection; *Chi-square* [16] assesses how selectively words and phrases co-occur within the same sentences as a particular subset of frequent terms in the document text. This 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; *RAKE* [20] hypothesised that keywords usually consist of multiple words and do not contain punctuation or stop words. It uses word co-occurrence information to determine the keywords.

**Machine Learning.** Machine Learning (ML) systems are often designed for specific entity classes and thus integrate term extraction and term classification. Machine Learning systems use training data to learn features useful for term extraction and classification. But the avaibility of reliable training resources is one of the main problems. Some proposed ATE approaches use machine learning (ML) [4] [24] [17]. Although ML may also generate noise and silence. The main challenge is how to select a set of discriminating features that can be used for accurate recognition (and classification) of term instances.

**Hybrid Methods.** Most approaches combine several methods (typically linguistic and statistically based) for the term extraction task. *GlossEx* [12] considers the probability of a word in the domain corpus divided by the probability of the appearance 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 differs from that in a general corpus. *C/NC-value* [6] combines statistical and linguistic information for the extraction of multiword and nested terms. This is the most well-known measure in the literature. While most studies address specific types of entities, *C/NC-value* is a domain-independent method. It has also been used for recognizing terms in the biomedical literature [8] [14]. In [25], the authors showed that *C-value* obtains the best results compared to the other measures cited above. Another measure is *F-TFIDF-C* [13], which combines an ATE measure *(C-value)* and an AKE measure *(TF-IDF)* to extract terms, thus

obtaining better results than *C-value*. Moreover, *C-value* has also been applied to different languages other than English, e.g., Japanese, Serbian, Slovenian, Polish, Chinese [10], Spanish [2], Arabic, and French [14]. That is why we have chosen *C-value* and *F-TFIDF-C* as baselines for the proposed experiments.

### 2.2 Graph-Based Approaches

Graph modeling is an alternative for modeling information, which clearly highlights relationships of nodes among vertices. It also groups related information in a specific way, and a centrality algorithm can be applied to enhance their efficiency. An increasingly popular recent application of graph approaches to Information Retrieval (IR) concerns social or collaborative networks and recommender systems [18]. Graph representations of text and scoring function definition are two widely explored research topics, but few studies have been focused on graph-based IR in terms of both document representation and weighting models [21]. First, text is modeled as a graph where nodes represent words and edges represent relations between words, defined on the basis of any meaningful statistical or linguistic relation [3]. In [3], the authors developed a graph-based word weighting model that represents each document as a graph. The importance of a word within a document is estimated by the number of related words and their importance, in the same way that PageRank [19] estimates the importance of a page via the pages that are linked to it. Another study, [21], introduces a different representation of document that captures relationships between words by using an unweighted directed graph of words with a novel scoring function.

In the above approaches, graphs are used to measure the influence of words in documents like automatic keyword extraction methods (AKE) while ranking documents against queries. These approaches differ from ours as they use graphs that are focused on the extraction of relevant words in a document and computing relations between words. In our proposal, a graph is built such that the vertices are multiword terms and the edges are relations between multiword terms. Moreover, we focus especially on a scoring function of relevant multiword terms in a domain rather than in a document.

## 3 Two Measures for Multiword Term Extraction

### 3.1 A New Ranking Measure Based on Linguistic and Statistical Information: *LIDF-value* (Linguisitic Patterns, IDF, and C-value Information)

Three steps are involved in computing the *LIDF-value*:

(1) **Part-of-Speech tagging:** a part-of-speech is applied to the whole corpus to obtain the lemma of words and to extract linguistic patterns. 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 in which it appears.

(2) **Candidate term extraction:** before applying any measures, we select terms having a syntactic structure appearing in the pattern list.
(3) **Ranking of candidate terms:** finally the *LIDF-value* is computed for each term.

These steps are explained in the next subsections and detailed in Algorithm 1.

**From the Linguistic-Based Approach.** The objective is to give greater importance to the term unithood in order to detect low frequency terms.

As in related work, we supposed that terms of a domain have a 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 work, we chose UMLS[1] which is a biomedical dictionary. We conduct part-of-speech tagging of the domain dictionary using the Stanford CoreNLP API (POS tagging)[2], and then compute the frequency of syntactic structures. Patterns among the 200 highest frequencies are selected to build the list. From this list, we compute the weight associated with the probability that a candidate term could be a domain term if its syntactic structure appears in the linguistic pattern list. In our experiments, 2 300 000 terms were used to build the list of patterns. Table 1 illustrates the computation of the linguistic pattern probability.

**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)

| 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 |

**To the Statistical-Based Approach.** Our method *LIDF-value* is aimed at computing the termhood for each term, using the *probability* calculated as defined above, the *idf*, and the *C-value* of each term. The inverse document frequency *(idf)* is a measure indicating the extent to which a 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 low frequency terms. The *C-value* measure is based on the term frequency. The aim of the *C-value* (see (1)) is to improve the extraction of nested terms, i.e., this criteria favors a candidate term that does not often appear in a longer term. For instance, in a specialized corpus (Ophthalmology), the authors of [6] found the irrelevant term "soft contact" while the frequent and longer term "soft contact lens" is relevant.

[1] `http://www.nlm.nih.gov/research/umls`
[2] `http://nlp.stanford.edu/software/corenlp.shtml`

$$C\text{-}value(A) = \begin{cases} \log_2(|A|) \times f(A) & \text{if } A \notin \text{ nested} \\ \log_2(|A|) \times \left( f(A) - \frac{1}{|S_A|} \times \sum_{b \in S_A} f(b) \right) & \text{otherwise} \end{cases} \quad (1)$$

Where $A$ represents multiword terms, $|A|$ the number of words in $A$, $f(A)$ the frequency of $A$ in the documents, $S_A$ the set of terms that contain $A$ and $|S_A|$ the number of terms in $S_A$. In a nutshell, *C-value* uses the frequency of the term if the term is not included in other terms (first line), or decreases this frequency if the term appears in other terms, by using the frequency of those other terms (second line). The algorithm 1 describes the applied process.

These different statistical information items (i.e., *probability* of linguisitic patterns, *C-value*, *idf*) are combined to define the global ranking measure *LIDF-value* (see (2)); where $\mathrm{P}(A_{LP})$ is the probability of a multiword term $A$ which has the same linguistic structure pattern $LP$, i.e., the weight of the linguistic pattern $LP$ computed in Sect. 3.1.

$$LIDF\text{-}value(A) = \mathrm{P}(A_{LP}) \times idf(A) \times C\text{-}value(A) \quad (2)$$

**Algorithm 1.** ComputeLIDF-value *(Corpus, Patterns, $min_{freq}$, $num_{terms}$)*

**Data**: *Corpus* = set of documents of a specific-domain;
*Patterns* = $HT_{patterns}$(*pattern, probability*) //Hashtable of linguistic patterns with its probability;
$min_{freq}$ = frequency threshold for candidate terms;
$num_{terms}$ = number of terms to take as output
**Result**: $L_{terms}$ = List of ranked terms
**begin**
  Tag the *Corpus*;
  Take the *lemma* of each tagged word;
  Extract candidate terms $A$ by filtering with *Patterns*;
  Remove candidate terms $A$ below $min_{freq}$;
  **for** *each candidate term* $A \in$ *Corpus* **do**
    $LIDF\text{-}value(A) = \mathrm{P}(A_{LP}) \times idf(A) \times C\text{-}value(A)$;
    add $A$ to $L_{terms}$;
  **end**
  Rank the $L_{terms}$ by the value obtained with *LIDF-value*;
  Select the first $num_{terms}$ terms of $L_{terms}$ ;
**end**

As an improvement, we propose to take into account graph-theoretic information to highlight relevant terms, as explained in the following subsection.

### 3.2 A New Graph-Based Ranking Measure: *TeRGraph* (Terminology Ranking Based on Graph Information)

This approach aims to improve the precision of the top $k$ extracted terms. As mentioned above, in contrast to the work cited before, the graph is built with a list of terms obtained according to the steps described in Sect. 3.1, where vertices denote multiword terms linked by their co-occurrence in the sentences in the corpus. Moreover, we apply the hypothesis that the term representativeness in a graph, for a specific-domain, depends on the number of neighbors that it has, and the number of neighbors of its neighbors. We assume that a term with more neighbors is less representative of the specific-domain. This means that this term is used in the general domain. Figure 1 illustrates our hypothesis. The graph-based approach is divided into two steps:


**Fig. 1.** Importance of a term in a domain

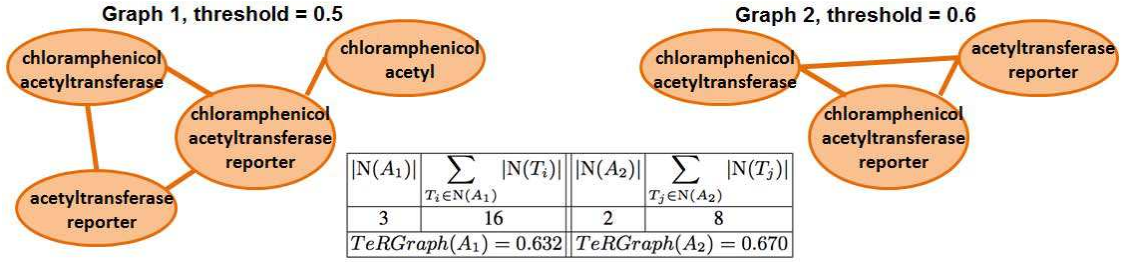
(1) **Graph construction:** a graph (see Fig. 2) is built where vertices denote terms, and edges denote co-occurrence relations between terms, co-occurrences between terms are measured as the weight of the relation in the initial corpus. This approach is statistical because it links all co-occurring terms without considering their meaning or function in the text. This graph is undirected as the edges imply that terms simply co-occur, without any further distinction regarding their role. We take *Dice coefficient*, a basic measure to compute the co-occurrence between two terms $x$ and $y$, defined by the following formula:

$$D(x, y) = \frac{2 \times P(x, y)}{P(x) + P(y)} \tag{3}$$

(2) **Representativeness computations on the term graph:** a principled graph-based measure to compute term weights (representativeness) is defined. The aim of this new graph ranking measure, *TeRGraph*, see (4), is to derive these weights for each vertex, (i.e., multiword term weight), in order to re-rank the list of extracted terms.

$$TeRGraph(A) = \log_2 \left( 1.5 + \frac{1}{|\mathrm{N}(A)| + \sum_{T_i \in \mathrm{N}(A)} |\mathrm{N}(T_i)|} \right) \quad (4)$$

Where $A$ represents a vertex (multiword term), $\mathrm{N}(A)$ the neighborhood of $A$, $|\mathrm{N}(A)|$ the number of neighbors of $A$, $T_i$ the neighbor $i$ of $A$. The intuition for (4) is as follows: the more a term $A$ has neighbors (directly with $\mathrm{N}(A)$ or by transitivity with $\mathrm{N}(T_i)$), the more the weight decreases. Indeed, a term $A$ having a lot of neighbors is considered too general for the domain (i.e., this term is not salient), then it has to be penalized via the associated score. Figure 2 shows an example to calculate the value of *TeRGraph* for a term in different graphs. These graphs are built with different co-occurrence thresholds (i.e., Dice's value between two terms). In this example, $A_1$ and $A_2$ represent the term *chloramphenicol acetyltransferase reporter* in Graphs 1 and 2 respectively.




| $\lvert\mathrm{N}(A_1)\rvert$ | $\sum_{T_i \in \mathrm{N}(A_1)} \lvert\mathrm{N}(T_i)\rvert$ | $\lvert\mathrm{N}(A_2)\rvert$ | $\sum_{T_j \in \mathrm{N}(A_2)} \lvert\mathrm{N}(T_j)\rvert$ |
|---|---|---|---|
| 3 | 16 | 2 | 8 |
| $TeRGraph(A_1) = 0.632$ | | $TeRGraph(A_2) = 0.670$ | |

**Fig. 2.** *TeRGraph*'s value for ***chloramphenicol acetyltransferase reporter***

## 4 Experiments and Results

### 4.1 Data, Protocol, and Validation

In our experiments, the standard GENIA[3] corpus was used, which is made up of 2 000 titles and abstracts of journal articles derived 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.

### 4.2 Results

The results are evaluated in terms of *precision* obtained over the top $k$ extracted terms (P@$k$) for the two proposed measures and baseline measures for multiword terms. In the following subsections, we narrow down the presented results by keepingfor the graph-based measureonly the first 8 000 extracted terms.

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

**Linguistic and Statistical Results.** Table 2 presents and compares the results of multiword term extraction with the best baseline measures, such as, *C-value*, *F-TFIDF-C*, and our measure *LIDF-value*. The best results were obtained with *LIDF-value* with an improvement in precision of 11% for the first hundred extracted multiword terms. The precision of *LIDF-value* will be further improved with *TeRGraph*.

**Table 2.** Precision comparison of *LIDF-value* with baseline measures

| | *C-value* | *F-TFIDF-C* | ***LIDF-value*** |
|---|---|---|---|
| P@100 | 0.690 | 0.715 | **0.820** |
| P@200 | 0.690 | 0.715 | **0.770** |
| P@300 | 0.697 | 0.710 | **0.750** |
| P@400 | 0.665 | 0.690 | **0.738** |
| P@500 | 0.642 | 0.678 | **0.718** |
| P@600 | 0.638 | 0.668 | **0.723** |
| P@700 | 0.627 | 0.669 | **0.717** |
| P@800 | 0.611 | 0.650 | **0.710** |
| P@900 | 0.612 | 0.629 | **0.714** |
| P@1000 | 0.605 | 0.618 | **0.697** |
| P@2000 | 0.570 | 0.557 | **0.662** |
| P@5000 | 0.498 | 0.482 | **0.575** |
| P@10000 | 0.428 | 0.412 | **0.526** |
| P@20000 | 0.353 | 0.314 | **0.377** |

We evaluated *LIDF-value* and baseline measures within a sequence of $n$-gram terms (i.e., $n$-gram term is a multiword term of $n$ words), for this we require an index term to be a $n$-gram terms of length $n \geq 2$. Table 3 shows the ranking of 3-gram terms with the baseline measures and *LIDF-value*. For 3-gram terms *C-value* obtains 2 irrelevant terms, *F-TFIDF-C* obtains 3 irrelevant terms while *LIDF-value* obtains only 1 irrelevant term.

**Graph Results.** Our graph-based approach is applied to the first 8 000 terms extracted by the Linguistic and Statistical approach. The objective is to re-rank the 8 000 terms while trying to improve the precision by intervals. One parameter is involved in the computation of graph-based term weights, namely the *threshold* of Dice value which represents the relation when building the term graph. This involves linking terms whose *Dice value* of the relation is higher than *threshold*. We vary *threshold* ($\delta$) within $\delta = [0.25, 0.35, 0.50, 0.60, 0.70]$ and report the precision performance for each of these values. Table 4 gives the precision performance obtained by *TeRGraph* and shows that it is well adapted for ATE.

**Summary.** Table 5 presents a precision comparison of our two measures. In terms of overall precision, our experiments produce consistent results from the GENIA corpus. In most cases, *TeRGraph* obtains better precision with a *threshold* of 0.60 and 0.70 (i.e., better precision in most P@$k$ intervals), which is very

**Table 3.** Comparison of top-10 ranked 3 gram terms (irrelevant terms are italicized and marked with *)

| *C-value* | *F-TFIDF-C* |
|---|---|
| human immunodeficiency virus | *kappa b alpha** |
| *kappa b alpha** | nf kappa b |
| tumor necrosis factor | jurkat t cell |
| electrophoretic mobility shift | human t cell |
| nf-kappa b activation | mhc class ii |
| *virus type 1** | cd4+ t cell |
| protein kinase c | *c-fos and c-jun** |
| long terminal repeat | peripheral blood monocyte |
| nf kappa b | t cell proliferation |
| jurkat t cell | *transcription factor nf-kappa** |
| ***LIDF-value*** | |
| i kappa b | |
| human immunodeficiency virus | |
| electrophoretic mobility shift | |
| human t cell | |
| mobility shift assay | |
| *kappa b alpha** | |
| tumor necrosis factor | |
| nf-kappa b activation | |
| protein kinase c | |
| jurkat t cell | |

**Table 4.** Precision performance of *TeRGraph* when varying $\delta$ parameter

| | $\delta \geq 0.25$ | $\delta \geq 0.35$ | $\delta \geq 0.50$ | $\delta \geq 0.60$ | $\delta \geq 0.70$ |
|---|---|---|---|---|---|
| P@100 | 0.840 | 0.860 | 0.910 | **0.930** | 0.900 |
| P@200 | 0.800 | 0.790 | 0.850 | **0.855** | **0.855** |
| P@300 | 0.803 | 0.773 | 0.833 | **0.830** | 0.820 |
| P@400 | 0.780 | 0.732 | **0.820** | **0.820** | 0.815 |
| P@500 | 0.774 | 0.712 | 0.798 | **0.810** | 0.806 |
| P@600 | 0.773 | 0.675 | 0.797 | **0.807** | 0.792 |
| P@700 | 0.760 | 0.647 | 0.769 | **0.796** | 0.787 |
| P@800 | 0.756 | 0.619 | 0.748 | **0.784** | 0.779 |
| P@900 | 0.748 | 0.584 | 0.724 | 0.773 | **0.777** |
| P@1000 | 0.751 | 0.578 | 0.720 | 0.766 | **0.769** |
| P@2000 | 0.689 | 0.476 | 0.601 | 0.657 | **0.694** |
| P@3000 | 0.642 | 0.522 | 0.535 | 0.605 | **0.644** |
| P@4000 | **0.612** | 0.540 | 0.543 | 0.559 | 0.593 |
| P@5000 | **0.574** | 0.546 | 0.544 | 0.554 | 0.562 |
| P@6000 | 0.558 | 0.539 | 0.540 | 0.549 | **0.561** |
| P@7000 | **0.556** | 0.540 | 0.540 | 0.545 | 0.552 |
| P@8000 | **0.546** | **0.546** | **0.546** | **0.546** | **0.546** |

**Table 5.** Precision comparison of *LIDF-value* and *TeRGraph*

| | *LIDF-value* | *TeRGraph* ($\delta \geq 0.60$) | *TeRGraph* ($\delta \geq 0.70$) |
|---|---|---|---|
| P@100 | 0.820 | **0.930** | 0.900 |
| P@200 | 0.770 | **0.855** | **0.855** |
| P@300 | 0.750 | **0.830** | 0.820 |
| P@400 | 0.738 | **0.820** | 0.815 |
| P@500 | 0.718 | **0.810** | 0.806 |
| P@600 | 0.723 | **0.807** | 0.792 |
| P@700 | 0.717 | **0.796** | 0.787 |
| P@800 | 0.710 | **0.784** | 0.779 |
| P@900 | 0.714 | 0.773 | **0.777** |
| P@1000 | 0.697 | 0.766 | **0.769** |
| P@2000 | 0.662 | 0.657 | **0.694** |
| P@3000 | 0.627 | 0.605 | **0.644** |
| P@4000 | **0.608** | 0.5585 | 0.593 |
| P@5000 | **0.575** | 0.5538 | 0.562 |
| P@6000 | 0.550 | 0.549 | **0.561** |
| P@7000 | 0.547 | 0.545 | **0.552** |
| P@8000 | **0.546** | **0.546** | **0.546** |

good because it helps alleviate the problem of manual validation of candidate terms. The performance of our graph-based measure depends somewhat on the value of the co-occurrence relation between terms. Specifically, the value of the co-occurrence relation affects how the graph is built (whose edges are taken), and hence it is critical for computation of the graph-based term weight. Another performance factor of our graph-based measure is the quality of the results obtained with *LIDF-value* due to the fact that to re-rank *TeRGraph* the list of terms extracted with *LIDF-value* is required as input, in order to construct the graph, where nodes denote terms, and edges denote co-occurrence relations.

## 5 Conclusions and Future Work

This paper defines and evaluates two measures for automatic multiword term extraction. The first one, *LIDF-value*, a linguistic and statistical-based measure, improves the precision of automatic term extraction in comparison with the most popular term extraction measure. This measure overcomes the lack of frequency information with the values of *linguistic pattern probability* and *idf*. We experimentally show that *LIDF-value* applied in the biomedical domain outperformed a state-of-the-art baseline for extracting terms (i.e., *C-value* and *F-TFIDF-C*), while obtaining the best precision results in all intervals (i.e., P@$k$).

The second one, *TeRGraph*, is a graph-based measure. It enables a reduction in the huge human effort required to validate candidate terms. The graph-based measure has never been applied for automatic term extraction. *TeRGraph* takes into account the neighborhood to compute the term representativeness in a

specific domain. Our experimental evaluations reveal that *TeRGraph* has better precision than *LIDF-value* for all intervals.

As a future extension of this work, we intend to use the relation value within *TeRGraph*. Moreover, we plan to test this general approach in other domains, such as ecology and agronomy. Finally, future work includes the use of other graph ranking computations, e.g., PageRank, adapted for automatic term extraction.

**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 of Montpellier 2 and CNRS.

## References

1. Ahmad, K., Gillam, L., Tostevin, L.: University of Surrey Participation in TREC8: Weirdness Indexing for Logical Document Extrapolation, Retrieval (WILDER). In: TREC (1999)
2. Barrón-Cedeño, A., Sierra, G., Drouin, P., Ananiadou, S.: An improved automatic term recognition method for Spanish. In: Gelbukh, A. (ed.) CICLing 2009. LNCS, vol. 5449, pp. 125–136. Springer, Heidelberg (2009)
3. Blanco, R., Lioma, C.: Graph-based term weighting for information retrieval. Information Retrieval 15, 54–92 (2012)
4. Conrado, M.S., Pardo, T.A.S., Rezende, S.O.: Exploration of a Rich Feature Set for Automatic Term Extraction. In: Castro, F., Gelbukh, A., González, M. (eds.) MICAI 2013, Part I. LNCS (LNAI), vol. 8265, pp. 342–354. Springer, Heidelberg (2013)
5. Dobrov, B., Loukachevitch, N.: Multiple Evidence for Term Extraction in Broad Domains. In: Proceeding of Recent Advances in Natural Language Processing (RANLP), Hissar, Bulgaria, pp. 710–715 (2011)
6. Frantzi, K., Ananiadou, S., Mima, H.: Automatic recognition of multiword terms: the C-value/NC-value Method. International Journal on Digital Libraries 3, 115–130 (2000)
7. Gaizauskas, R., Demetriou, G., Humphreys, K.: Term recognition, classification in biological science journal articles. In: Proceeding of the Computional Terminology for Medical, Biological Applications Workshop of the 2nd International Conference on NLP, pp. 37–44 (2000)
8. Hliaoutakis, A., Zervanou, K., Petrakis, E.G.M.: The AMTEx approach in the medical document indexing, retrieval application. Data & Knowl. Engineering 68, 380–392 (2009)
9. Ittoo, A., Bouma, G.: Term Extraction from Sparse, Ungrammatical Domain-specific Documents. Expert Systems with Applications 40, 2530–2540 (2013)
10. Ji, L., Sum, M., Lu, Q., Li, W., Chen, Y.: Chinese Terminology Extraction Using Window-Based Contextual Information. In: Gelbukh, A. (ed.) CICLing 2007. LNCS, vol. 4394, pp. 62–74. Springer, Heidelberg (2007)
11. Kageura, K., Umino, B.: Methods of automatic term recognition: A review. Terminology 3, 259–289 (1996)

12. Kozakov, L., Park, Y., Fin, T., Drissi, Y., Doganata, N., Confino, T.: Glossary extraction, knowledge in large organisations via semantic web technologies. In: Proceedings of the 6th International Semantic Web Conference, he 2nd Asian Semantic Web Conference (Semantic Web Challenge Track) (2004)
13. Lossio-Ventura, J.A., Jonquet, C., Roche, M., Teisseire, M.: Biomedical Terminology Extraction: A new combination of Statistical, Web Mining Approaches. In: Proceedings of Journées Internationales d'Analyse Statistique des Données Textuelles (JADT 2014), Paris, France (2014)
14. Lossio-Ventura, J.A., Jonquet, C., Roche, M., Teisseire, M.: Combining C-value, Keyword Extraction Methods for Biomedical Terms Extraction. In: Proceedings of the Fifth International Symposium on Languages in Biology, Medicine (LBM 2013), Tokyo, Japan, pp. 45–49 (2013)
15. Lossio-Ventura, J.A., Hacid, H., Ansiaux, A., Maag, M.L.: Conversations reconstruction in the social web. In: Proceedings of the 21st International Conference Companion on World Wide Web (WWW 2012), pp. 573–574. ACM, Lyon (2012)
16. Matsuo, Y., Ishizuka, M.: Keyword extraction from a single document using word co-occurrence statistical information. International Journal on Artificial Intelligence Tools 13, 157–169 (2004)
17. Newman, D., Koilada, N., Lau, J.H., Baldwin, T.: Bayesian Text Segmentation for Index Term Identification, Keyphrase Extraction. In: Proceedings of 24th International Conference on Computational Linguistics, Mumbai, India, pp. 2077–2092 (2012)
18. Noh, T., Park, S., Yoon, H., Lee, S., Park, S.: An Automatic Translation of Tags for Multimedia Contents Using Folksonomy Networks. In: Proceedings of the 32Nd International ACM SIGIR Conference on Research, Development in Information Retrieval, SIGIR 2009, pp. 492–499. ACM, Boston (2009)
19. Page, L., Brin, S., Motwani, R., Winograd, T.: The PageRank citation ranking: Bringing order to the web. Stanford InfoLab (1999)
20. Rose, S., Engel, D., Cramer, N., Cowley, W.: Automatic keyword extraction from individual documents. Text Mining: Theory, Applications, pp. 1–20. John Wiley, Sons, Ltd. (2010)
21. Rousseau, F., Vazirgiannis, M.: Graph-of-word, TW-IDF: New Approach to Ad Hoc IR. In: Proceedings of the 22nd ACM International Conference on Conference on Information, Knowledge Management, CIKM 2013, pp. 59–68. ACM, San Francisco (2013)
22. Stoykova, V., Petkova, E.: Automatic extraction of mathematical terms for precalculus. Procedia Technology Journal 1, 464–468 (2012)
23. Van Eck, N.J., Waltman, L., Noyons, E.C.M., Buter, R.K.: Automatic term identification for bibliometric mapping. Scientometrics 82, 581–596 (2010)
24. Zhang, X., Song, Y., Fang, A.C.: Term recognition using conditional random fields. In: International Conference on Natural Language Processing, Knowledge Engineering (NLP-KE), pp. 1–6. IEEE (2010)
25. Zhang, Z., Iria, J., Brewster, C., Ciravegna, F.: A Comparative Evaluation of Term Recognition Algorithms. In: Proceedings of the Sixth International Conference on Language Resources, Evaluation (LREC 2008), Marrakech, Morocco (2008)