HAL
archives-ouvertes.fr

Extraction automatique de termes combinant différentes informations

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

▶ **To cite this version:**

Juan Antonio Lossio-Ventura, Clement Jonquet, Mathieu Roche, Maguelonne Teisseire. Extraction automatique de termes combinant différentes informations. TALN: Traitement Automatique des Langues Naturelles, Jul 2014, Marseille, France. 21ème, pp.407-412, 2014, <http://www.taln2014.org/site/>. <lirmm-01020051>

HAL Id: lirmm-01020051
https://hal-lirmm.ccsd.cnrs.fr/lirmm-01020051

Submitted on 7 Jul 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.

Extraction automatique de termes combinant différentes informations

Juan Antonio Lossio-Ventura[1] Clement Jonquet[1] Mathieu Roche[1,2] Maguelonne Teisseire[1,2]
(1) LIRMM, Université de Montpellier 2, CNRS, Montpellier - France
(2) Irstea, Cirad, TETIS, Montpellier - France
juan.lossio@lirmm.fr, clement.jonquet@lirmm.fr, mathieu.roche@cirad.fr,
teisseire@teledetection.fr

Résumé. Pour une communauté, la terminologie est essentielle car elle permet de décrire, échanger et récupérer les données. Dans de nombreux domaines, l'explosion du volume des données textuelles nécessite de recourir à une automatisation du processus d'extraction de la terminologie, voire son enrichissement. L'extraction automatique de termes peut s'appuyer sur des approches de traitement du langage naturel. Des méthodes prenant en compte les aspects linguistiques et statistiques proposées dans la littérature, résolvent quelques problèmes liés à l'extraction de termes tels que la faible fréquence, la complexité d'extraction de termes de plusieurs mots, ou l'effort humain pour valider les termes candidats. Dans ce contexte, nous proposons deux nouvelles mesures pour l'extraction et le "ranking" des termes formés de plusieurs mots à partir des corpus spécifiques d'un domaine. En outre, nous montrons comment l'utilisation du Web pour évaluer l'importance d'un terme candidat permet d'améliorer les résultats en terme de précision. Ces expérimentations sont réalisées sur le corpus biomédical GENIA en utilisant des mesures de la littérature telles que *C-value*.

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, rely 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*.

Mots-clés : Extraction Automatique de Termes, Mesure basée sur le Web, Mesure Linguistique, Mesure Statistique, Traitement Automatique du Langage Biomédical.

Keywords: Automatic Term Extraction, Web-based measure, Linguistic-based measure, Statistic-based measure, Biomedical Natural Language Processing.

1 Introduction

Les méthodes d'Extraction Automatique de Termes (EAT) visent à extraire automatiquement des termes techniques à partir d'un corpus. Ces méthodes sont essentielles pour l'acquisition des connaissances d'un domaine pour des tâches telles que la mise à jour de lexique. En effet, les termes techniques sont importants pour mieux comprendre le contenu d'un domaine. Ces termes peuvent être : (i) composés d'un seul mot (généralement simple à extraire), ou (ii) composés de plusieurs mots (difficile à extraire). Notre travail concerne plus spécifiquement l'extraction de termes composés de plusieurs mots.

Les méthodes d'EAT impliquent généralement deux étapes principales. La première extrait des candidats en calculant "l'unithood" qui qualifie une chaîne de mots comme une expression valide (Korkontzelos *et al.*, 2008). La deuxième étape calcule le "termhood" qui sert à mesurer la spécificité propre à un domaine.

Il existe quelques problèmes connus de l'EAT tels que : (i) l'extraction de termes non pertinents (bruit) ou le nombre réduit de termes pertinents retournés (silence), (ii) l'extraction de termes de plusieurs mots qui ont inévitablement des structures complexes, (iii) l'effort humain dans la validation manuelle des termes candidats, (iv) l'application aux corpus

de grande échelle. En réponse à ces problèmes, nous proposons deux nouvelles mesures. La première, appelée *LIDF-value*, est fondée sur l'information statistique et linguistique. La mesure *LIDF-value* permet de mieux prendre en compte l'unithood, en lui adossant un niveau de qualité. Elle traite les problèmes i), ii) et iv). La seconde, appelée *WAHI*, est une mesure fondée sur le Web traitant les problèmes i), ii) et iii). Dans cet article, nous comparons la qualité des méthodes proposées avec les mesures de référence les plus utilisées. Nous démontrons que l'utilisation de ces deux mesures améliore l'extraction automatique de termes spécifiques d'un domaine, à partir de textes qui n'offrent pas une fiabilité statistique liée aux fréquences.

Le reste du papier est organisé comme suit : nous discutons tout d'abord des travaux connexes dans la Section 2. Les deux nouvelles mesures sont ensuite décrites en Section 3. L'évaluation en termes de précision est présentée dans la Section 4 suivie des conclusions en Section 5.

2 État de l'art

Plusieurs études récentes se sont concentrées sur l'extraction des termes de plusieurs mots (n-grammes) et d'un seul mot (unigrammes). Les méthodes d'extraction de terme existantes peuvent être divisées en quatre grandes catégories : (i) *linguistique*, (ii) *statistique*, (iii) *apprentissage automatique*, et (iv) *hybride*. La plupart de ces techniques appartiennent à des approches de fouille de textes. Les techniques existantes fondées sur le Web ont rarement été appliquées à l'EAT, mais comme nous le verrons, ces approches peuvent être adaptées à cet objectif.

Méthodes de Fouille de Textes

Les méthodes liées à la fouille de textes combinent en général différents types d'approches : linguistique (Gaizauskas *et al.*, 2000; Krauthammer & Nenadic, 2004), statistiques (Van Eck *et al.*, 2010) ou fondées sur un apprentissage automatique pour l'extraction et la classification des termes (Newman *et al.*, 2012). Il faut également citer les propositions issues du domaine de l'Extraction Automatique de Mots Clés (EAMC) dont les mesures peuvent être adaptées pour l'extraction de termes d'un corpus (Lossio-Ventura *et al.*, 2013) (Lossio-Ventura *et al.*, 2014).

Les méthodes hybrides sont principalement linguistiques et statistiques. *GlossEx* (Kozakov *et al.*, 2004) estime la probabilité du mot dans un corpus de domaine comparée à la probabilité du même mot dans un corpus général. *Weirdness* (Ahmad *et al.*, 1999) estime que la distribution des mots dans un corpus spécifique est différente de la distribution des mots dans un corpus général. *C/NC-value* (Frantzi *et al.*, 2000), qui combine l'information statistique et linguistique pour l'extraction de termes de plusieurs mots et des termes imbriqués, est la mesure la plus connue dans le domaine biomédical. Dans (Zhang *et al.*, 2008), les auteurs montrent que *C-value* a des meilleurs résultats par rapport à d'autres mesures citées ci-dessus. Une autre mesure est *F-TFIDF-C* (Lossio-Ventura *et al.*, 2014) qui combine une mesure d'EAT *(C-value)* et une mesure d'EAMC *(TF-IDF)* pour extraire des termes obtenant des résultats plus satisfaisants que *C-value*. Aussi, *C-value* a été appliquée à de nombreuses langues autres que l'anglais, comme le japonais, serbe, slovène, polonais, chinois, espagnol, arabe, et français (Ji *et al.*, 2007; Barron-Cedeno *et al.*, 2009; Lossio-Ventura *et al.*, 2013). Ainsi, nous proposons d'adopter *C-value* et *F-TFIDF-C* comme référence pour l'étude comparative au cours de nos expérimentations.

Méthodes de Fouille du Web

Différentes études de fouille du Web se concentrent sur la similarité et les relations sémantiques. Les mesures d'association de mots peuvent être divisées en trois catégories (Chaudhari *et al.*, 2011) : (i) *Co-occurrence* qui s'appuient sur les fréquences de co-occurrence de deux mots dans un corpus, (ii) *Basées sur la similarité distributionnelle* qui caractérisent un mot par la distribution d'autres mots qui l'entourent, et (iii) *Basées sur les connaissances*, comme les thésaurus, les réseaux sémantiques, ou taxonomies. Dans cet article, nous nous concentrons sur les mesures de co-occurrence, car notre objectif est d'extraire les termes de plusieurs mots et nous proposons de calculer un degré d'association entre les mots qui composent un terme. Les mesures d'association de mots sont utilisées dans plusieurs domaines comme l'écologie, la psychologie, la médecine et le traitement du langage. De telles mesures ont été récemment étudiées dans (Pantel *et al.*, 2009) (Zadeh & Goel, 2013), telles que *Dice*, *Jaccard*, *Overlap*, *Cosine*. Une autre mesure pour calculer l'association entre les mots utilisant les résultats des moteurs de recherche Web est la Normalized Google Distance (Cilibrasi & Vitanyi, 2007). Elle s'appuie sur le nombre de fois où les mots apparaissent ensemble dans le document indexé par un moteur de recherche. Dans cette étude, nous comparons les résultats de notre mesure fondée sur le Web avec les mesures d'association de référence (*Dice*, *Jaccard*, *Overlap*, *Cosine*).

3 Deux nouvelles mesures pour l'extraction de termes

3.1 Une mesure fondée sur l'information linguistique et statistique : *LIDF-value*

Notre première contribution consiste à donner une meilleure importance à l'unithood des termes afin de détecter des termes avec une faible fréquence.

De manière similaire aux travaux de la littérature, nous supposons que les termes d'un domaine ont une structure syntaxique similaire. Par conséquent, nous construisons une liste de patrons linguistiques les plus courants selon la structure syntaxique des termes techniques présents dans un dictionnaire. Dans notre cas, il s'agit d'UMLS [1] qui est un ensemble de référence de terminologies biomédicales. Dans un premier temps, nous effectuons l'étiquetage grammatical des termes contenus dans le dictionnaire en utilisant Stanford CoreNLP API (POS tagging) [2]. Ensuite nous calculons la fréquence des structures syntaxiques. Nous sélectionnons les 200 fréquences les plus élevées pour construire la liste de patrons (ou motifs) qui seront pris en considération. Cette liste est pondérée selon la fréquence d'apparition de chaque patron par rapport à l'ensemble des motifs. Un terme candidat est alors retenu s'il appartient à la liste des structures syntaxiques sélectionnées. Le nombre de termes utilisés pour construire cette liste était de 2 300 000. La Figure 1 illustre le calcul de la *probabilité* des patrons linguistiques.

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

FIGURE 1: Exemple de construction de patrons linguistiques (où *NN* : nom, *IN* : préposition, *JJ* : adjectif, et *CD* : numéro).

L'objectif de notre mesure, *LIDF-value* (***L**inguisitic patterns*, ***IDF***, and ***C-value*** information) est de calculer le termhood pour chaque terme, en utilisant la *probabilité* calculée précédemment, également avec l'idf, et *C-value* de chaque terme. La fréquence inverse de document (idf) est une mesure indiquant si le terme est commun ou rare dans tous les documents. Il est obtenu en divisant le nombre total de documents par le nombre de documents contenant le terme, puis en prenant le logarithme de ce quotient. La *probabilité* et l'idf améliorent la pondération des termes de faible fréquence.

En outre, la mesure *C-value* est fondée sur la fréquence des termes. Le but de *C-value* (Formule 1) est d'améliorer l'extraction de termes imbriqués. Ce critère favorise les termes candidats n'apparaissant pas dans des termes plus longs. Par exemple, dans un corpus spécialisé (ophtalmologie), (Frantzi *et al.*, 2000) ont trouvé le terme non pertinent *soft contact* cependant le terme plus long *soft contact lens* est pertinent.

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

Où A est un terme de plusieurs mots, $|A|$ le nombre de mots de A ; $f(A)$ la fréquence du terme A, S_A l'ensemble de termes qui contiennent A et $|S_A|$ le nombre de termes de S_A. Ainsi, *C-value* utilise soit la fréquence du terme si le terme n'est pas inclus dans d'autres termes (première ligne), ou diminue cette fréquence si le terme apparaît dans d'autres termes (deuxième ligne).

Nous avons combiné ces différentes informations statistiques (i.e., *probabilité* des patrons linguistiques, $C\text{-}value$, idf) pour proposer une nouvelle mesure globale de ranking appelée *LIDF-value* (Formule 2). Dans cette formule, A représente un terme de plusieurs mots ; $\mathrm{P}(A_{LP})$ la probabilité du patron linguistique LP associé au terme A qui a la même structure que le patron linguistique LP, c'est-à-dire le poids du patron linguistique LP calculé précédemment.

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

Ainsi, pour calculer *LIDF-value* nous exécutons trois étapes, résumées ci-dessous :

(1) **Étiquetage grammatical :** nous effectuons l'étiquetage morpho-syntaxique du corpus, ensuite nous considérons le lemme de chaque mot.
(2) **Extraction de termes candidats :** avant d'appliquer les mesures, nous filtrons notre corpus en utilisant les patrons calculés précédemment. Nous choisissons uniquement les termes qui ont une structure syntaxique présente dans la liste de patrons sélectionnés.
(3) **Ranking de termes candidats :** enfin, nous calculons la valeur de *LIDF-value* pour chaque terme.

Afin d'améliorer le classement des termes pertinents, nous proposons, dans la sous-section suivante, de prendre en compte l'information Web.

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

3.2 Une nouvelle mesure de ranking fondée sur le Web : *WAHI*

Des travaux associés à la fouille du Web interrogent les moteurs de recherche pour mesurer l'association entre les mots. Ceci peut être utilisé pour mesurer l'association des mots qui composent un terme (e.g., *soft*, *contact*, et *lens* qui composent le terme pertinent *soft contact lens*). Dans nos travaux, nous proposons d'associer le critère de Dice avec une mesure d'association appelée $WebR$ (Lossio-Ventura *et al.*, 2014) (Formule 3). Par exemple pour le terme *soft contact lens*, le numérateur correspond au nombre de pages Web avec la requête *"soft contact lens"*, et le dénominateur correspond au résultat de la requête *soft ET contact ET lens*.

$$WebR(A) = \frac{nb(\text{“}A\text{”})}{nb(A)} \tag{3}$$

La mesure $WAHI$ (**W**eb **A**ssociation based on **H**its **I**nformation) que nous proposons, combine $Dice$ et $WebR$ de la manière suivante :

$$WAHI(A) = \frac{n \times nb(\text{“}A\text{”})}{\sum_{i=1}^{n} nb(a_i)} \times \frac{nb(\text{“}A\text{”})}{nb(A)} \tag{4}$$

Où a_i est un mot, $a_i \in A$ et $a_i = \{nom, adjectif, mot\ etranger\}$. Nous montrons que les ressources de domaine ouvert, telles que le Web, peuvent être exploitées pour aider l'extraction de termes spécifiques.

4 Expérimentations

4.1 Corpus et Protocole

Dans nos expérimentations, nous utilisons le corpus GENIA [3], qui est composé de 2 000 titres et des résumés d'articles des journaux issus de Medline, avec plus de 400 000 mots. GENIA contient des expressions linguistiques qui font référence à des entités d'intérêt en biologie moléculaire telles que les protéines, les gènes et les cellules. L'annotation des termes techniques couvre l'identification des entités physiques biologiques ainsi que d'autres termes importants. Afin de mettre en place un protocole de validation automatique et de couvrir les termes médicaux, nous créons un dictionnaire qui contient tous les termes d'UMLS ainsi que tous les termes techniques de GENIA. De cette manière nous pouvons évaluer la précision avec un dictionnaire de référence plus complet.

4.2 Résultats

Les résultats sont évalués en termes de *précision* obtenue sur les k premiers termes ($P@k$) pour les deux mesures présentées dans la section précédente.

Résultats de *LIDF-value* : Le tableau 1 compare les résultats de *C-value*, *F-TFIDF-C*, avec notre mesure *LIDF-value*. Le meilleur résultat est obtenu par *LIDF-value* pour l'ensemble des valeurs de k. *LIDF-value* est donc plus performante que les mesures de référence avec un gain en précision de 11 points pour les 100 premiers termes extraits. Ces résultats de précision sont illustrés dans la Figure 2.

	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 1: Précision selon le nombre de termes ($P@k$)

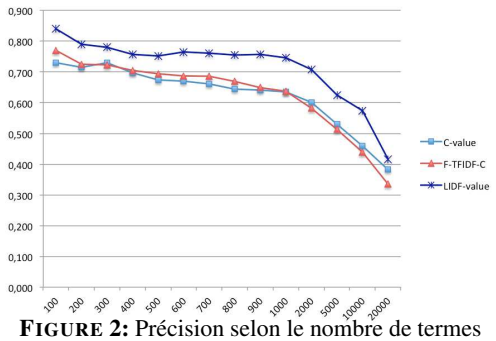

FIGURE 2: Précision selon le nombre de termes

Résultats des indexes de termes : Nous avons évalué *LIDF-value* et les mesures de référence avec une séquence de n-grammes de mots (i.e., un n-grammes de mots est un terme de n mots, par exemple *human immunodeficiency virus* est un 3-grammes de mots). Pour cela, nous construisons un index composé de n-grammes de mots ($n \geq 2$). Nous expérimentons la performance de *LIDF-value* sur les n-grammes de mots en prenant les 1 000 premiers termes. Le Tableau 2 présente la comparaison de la précision pour les 2-grammes, 3-grammes et 4+ grammes de mots.

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

	2-grammes de mots			*3-grammes de mots*			*4+ grammes de mots*		
	C-value	*F-TFIDF-C*	***LIDF-value***	*C-value*	*F-TFIDF-C*	***LIDF-value***	*C-value*	*F-TFIDF-C*	***LIDF-value***
P@100	0.770	0.760	**0.830**	0.670	0.530	**0.820**	0.510	0.370	**0.640**
P@200	0.755	0.755	**0.805**	0.590	0.450	**0.795**	0.455	0.330	**0.520**
P@300	0.710	0.743	**0.790**	0.577	0.430	**0.777**	0.387	0.273	**0.477**
P@400	0.695	0.725	**0.768**	0.560	0.425	**0.755**	0.393	0.270	**0.463**
P@500	0.692	0.736	**0.752**	0.548	0.398	**0.744**	0.378	0.266	**0.418**
P@600	0.683	0.733	**0.763**	0.520	0.378	**0.720**	0.348	0.253	**0.419**
P@700	0.670	0.714	**0.757**	0.499	0.370	**0.706**	0.346	0.249	**0.390**
P@800	0.669	0.703	**0.749**	0.488	0.379	**0.691**	0.323	0.248	**0.395**
P@900	0.654	0.692	**0.749**	0.482	0.399	**0.667**	0.323	0.240	**0.364**
P@1000	0.648	0.684	**0.743**	0.475	0.401	**0.660**	0.312	0.232	**0.354**

TABLE 2: Comparaison de précision des 2-grammes de mots, 3-grammes de mots et 4+ grammes de mots

Résultats de *WAHI* : Notre approche de fouille du Web est appliquée à la fin du processus, avec les 1 000 premiers termes extraits avec les mesures linguistiques et statistiques. La raison principale de cette limitation est le nombre restreint de requêtes autorisées par les moteurs de recherche. À cette étape, l'objectif est de faire le re-ranking des 1 000 termes améliorant la précision par intervalles. Le Tableau 3 montre la précision obtenue après le re-ranking avec *WAHI* et les mesures d'association de référence, utilisant les moteurs de recherche *Yahoo* et *Bing*. Ce tableau souligne que *WAHI* est bien adaptée pour l'EAT obtenant des meilleurs résultats de précision que les mesures de référence.

	YAHOO					*BING*				
	WAHI	*Dice*	*Jaccard*	*Cosine*	*Overlap*	***WAHI***	*Dice*	*Jaccard*	*Cosine*	*Overlap*
P@100	**0.900**	0.720	0.720	0.76	0.730	**0.800**	0.740	0.730	0.680	0.650
P@200	**0.800**	0.775	0.770	0.740	0.765	**0.800**	0.775	0.775	0.735	0.705
P@300	**0.800**	0.783	0.780	0.767	0.753	**0.800**	0.770	0.763	0.740	0.713
P@400	**0.800**	0.770	0.765	0.770	0.740	**0.800**	0.765	0.765	0.752	0.712
P@500	**0.820**	0.764	0.754	0.762	0.738	**0.800**	0.760	0.762	0.758	0.726
P@600	**0.767**	0.748	0.740	0.765	0.748	**0.817**	0.753	0.752	0.753	0.743
P@700	**0.786**	0.747	0.744	0.747	0.757	**0.814**	0.7514	0.751	0.733	0.749
P@800	**0.775**	0.752	0.7463	0.740	0.760	**0.775**	0.745	0.747	0.741	0.754
P@900	**0.756**	0.749	0.747	0.749	0.747	**0.778**	0.747	0.748	0.742	0.748
P@1000	**0.746**	**0.746**	**0.746**	**0.746**	**0.746**	**0.746**	**0.746**	**0.746**	**0.746**	**0.746**

TABLE 3: Comparaison de précision de *WAHI avec YAHOO et BING* et les mesures d'association

Discussion. *LIDF-value* obtient les meilleurs résultats de précision sur tous les intervalles pour l'extraction des termes et pour l'extraction de n-grammes de mots. Le tableau 4 présente les précisions obtenues par nos deux mesures sur le corpus GENIA. *WAHI* basée sur Yahoo obtient une meilleure précision (90 %) pour $P@100$. En comparaison, *WAHI* basée sur Bing obtient une précision de 80 %. Pour les autres intervalles, le Tableau 4 montre que *WAHI* fondée sur Bing obtient en général des résultats légèrement meilleurs. La performance de *WAHI* dépend du moteur de recherche adopté du fait de l'algorithme d'indexation associé. Enfin, le Tableau 4 montre que le re-ranking avec *WAHI* augmente la précision de *LIDF-value*.

	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 4: Précisions obtenues par *LIDF-value* et *WAHI* selon le nombre de terme ($P@k$)

5 Conclusions et Futurs travaux

L'article présente deux mesures pour l'extraction automatique de termes composés de plusieurs mots. La première est une mesure statistique et linguistique, *LIDF-value*, qui améliore la précision de l'extraction automatique de termes en comparaison avec les mesures classiques. Elle permet de compenser le manque d'information propre à la fréquence avec les valeurs des *probabilités des patrons linguistiques* et *idf*. La seconde, *WAHI*, est une mesure fondée sur le Web et prend comme entrée la liste de termes obtenus avec *LIDF-value*. Cette mesure permet de réduire l'effort humain conséquent

nécessaire à la validation des termes candidats. Nous montrons expérimentalement que *LIDF-value* offre des meilleurs résultats que les mesures de référence pour l'extraction de n-grammes de mots issus du domaine biomédical sur le corpus GENIA. Par ailleurs, ces résultats sont améliorés par l'utilisation de la mesure *WAHI*.

Les perspectives à ce travail sont nombreuses. Tout d'abord, nous souhaitons utiliser le Web pour extraire des termes plus longs que ceux actuellement obtenus. De plus, nous projetons de tester cette approche générale sur d'autres domaines, tels que l'écologie et l'agronomie. Enfin, nous visons à expérimenter notre proposition sur des corpus d'autres langues telles que le français et l'espagnol.

Références

AHMAD K., GILLAM L. & TOSTEVIN L. (1999). University of surrey participation in trec8 : Weirdness indexing for logical document extrapolation and retrieval (wilder). In *TREC*.

BARRON-CEDENO A., SIERRA G., DROUIN P. & ANANIADOU S. (2009). An improved automatic term recognition method for spanish. In *Computational Linguistics and Intelligent Text Processing*, p. 125–136. Springer.

CHAUDHARI D. L., DAMANI O. P. & LAXMAN S. (2011). Lexical co-occurrence, statistical significance, and word association. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, p. 1058–1068, Stroudsburg, PA, USA : Association for Computational Linguistics.

CILIBRASI R. L. & VITANYI P. M. (2007). The google similarity distance. *Knowledge and Data Engineering, IEEE Transactions on*, **19**(3), 370–383.

FRANTZI K., ANANIADOU S. & MIMA H. (2000). Automatic recognition of multi-word terms : the c-value/nc-value method. *International Journal on Digital Libraries*, **3**(2), 115–130.

GAIZAUSKAS R., DEMETRIOU G. & HUMPHREYS K. (2000). 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*, p. 37–44.

JI L., SUM M., LU Q., LI W. & CHEN Y. (2007). Chinese terminology extraction using window-based contextual information. In *Proceedings of the 8th International Conference on Computational Linguistics and Intelligent Text Processing (CICLing07)*, p. 62–74, Berlin, Heidelberg : Springer-Verlag.

KORKONTZELOS I., KLAPAFTIS I. P. & MANANDHAR S. (2008). Reviewing and evaluating automatic term recognition techniques. In *Advances in Natural Language Processing*, p. 248–259. Springer.

KOZAKOV L., PARK Y., FIN T., DRISSI Y., DOGANATA N. & CONFINO T. (2004). Glossary extraction and knowledge in large organisations via semantic web technologies. In *Proceedings of the 6th International Semantic Web Conference and he 2nd Asian Semantic Web Conference (Se-mantic Web Challenge Track)*.

KRAUTHAMMER M. & NENADIC G. (2004). Term identification in the biomedical literature. *Journal of Biomedical Informatics*, **37**(6), 512–526.

LOSSIO-VENTURA J. A., JONQUET C., ROCHE M. & TEISSEIRE M. (2013). Combining c-value and keyword extraction methods for biomedical terms extraction. In *Proceedings of the Fifth International Symposium on Languages in Biology and Medicine (LBM13)*, p. 45–49, Tokyo, Japan.

LOSSIO-VENTURA J. A., JONQUET C., ROCHE M. & TEISSEIRE M. (2014). 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.

NEWMAN D., KOILADA N., LAU J. H. & BALDWIN T. (2012). Bayesian text segmentation for index term identification and keyphrase extraction. In *Proceedings of 24th International Conference on Computational Linguistics (COLING)*, p. 2077–2092, Mumbai, India.

PANTEL P., CRESTAN E., BORKOVSKY A., POPESCU A.-M. & VYAS V. (2009). Web-scale distributional similarity and entity set expansion. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '09, p. 938–947, Stroudsburg, PA, USA : Association for Computational Linguistics.

VAN ECK N. J., WALTMAN L., NOYONS E. C. & BUTER R. K. (2010). Automatic term identification for bibliometric mapping. *Scientometrics*, **82**(3), 581–596.

ZADEH R. B. & GOEL A. (2013). Dimension independent similarity computation. *Journal of Machine Learning Research*, **14**(1), 1605–1626.

ZHANG Z., IRIA J., BREWSTER C. & CIRAVEGNA F. (2008). A Comparative Evaluation of Term Recognition Algorithms. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC08)*, Marrakech, Morocco.