Un critère de cohésion thématique fondé sur un graphe de cooccurrences

Clément de Groc[1, 2] Xavier Tannier[2, 3] Claude de Loupy[1]
(1) Syllabs, 15 rue Jean-Baptiste Berlier, 75013 Paris
(2) Univ. Paris-Sud, 91403 Orsay Cedex
(3) LIMSI-CNRS, B.P. 133, 91403 Orsay Cedex
cdegroc@limsi.fr, xtannier@limsi.fr, loupy@syllabs.com

RÉSUMÉ
Dans cet article, nous définissons un nouveau critère de cohésion thématique permettant de pondérer les termes d'un lexique thématique en fonction de leur pertinence. Le critère s'inspire des approches *Web as corpus* pour accumuler des connaissances exogènes sur un lexique. Ces connaissances sont ensuite modélisées sous forme de graphe et un algorithme de marche aléatoire est appliqué pour attribuer un score à chaque terme. Après avoir étudié les performances et la stabilité du critère proposé, nous l'évaluons sur une tâche d'aide à la création de lexiques bilingues.

ABSTRACT
Topical Cohesion using Graph Random Walks

In this article, we propose a novel metric to weight specialized lexicons terms according to their relevance to the underlying thematic. Our method is inspired by Web as corpus approaches and accumulates exogenous knowledge about a specialized lexicon from the web. Terms cooccurrences are modelled as a graph and a random walk algorithm is applied to compute terms relevance. Finally, we study the performance and stability of the metric and evaluate it in a bilingual lexicon creation context.

MOTS-CLÉS : Cohésion thématique, graphe de cooccurrences, marche aléatoire.

KEYWORDS: Thematic relevance, cooccurrence graph, random walk.

Actes de la conférence conjointe JEP-TALN-RECITAL 2012, volume 2: TALN, pages 183–195,
Grenoble, 4 au 8 juin 2012. ©2012 ATALA & AFCP

1 Introduction

Les lexiques et les terminologies sont des éléments essentiels du traitement automatique des langues. Ils sont utilisés dans une grande variété de tâches, allant de la catégorisation de textes à l'analyse d'opinions. Dans cet article, nous nous intéressons plus particulièrement aux lexiques dits thématiques ou spécialisés, c'est-à-dire composés de termes pertinents pour un domaine particulier. La Table 1 présente un extrait de lexique thématique sur le domaine de l'astronomie.

soleil	étoile	rayon gamma	étoile à neutron	masse solaire	...
planète	disque d'accrétion	naine blanche	proto-étoile	pulsar	...
astronome	quasar	astronomie	trou noir	neutron	...

TABLE 1 – Extrait de lexique thématique sur l'astronomie

La construction manuelle de tels lexiques est une tâche laborieuse et coûteuse. C'est pourquoi l'utilisation du Web ou de traducteurs automatiques comme appui pour la création de lexiques et de corpus spécialisés est une idée maintenant largement répandue (Baroni et Bernardini, 2004; Groc *et al.*, 2011; Kilgarriff et Grefenstette, 2003; Wan, 2009). Bien que l'utilité de telles approches ne soit plus à démontrer, une étape de validation manuelle reste requise.

Dans cet article, nous proposons un nouveau critère de *cohésion thématique* permettant de pondérer les termes d'un lexique thématique en fonction de leur pertinence pour le thème. Nous utilisons le Web comme source de corpus spécialisés sur les termes d'un lexique thématique. Nous modélisons ensuite les cooccurrences entre les termes du lexique sous la forme d'un graphe orienté où les sommets sont les termes du lexique et les arcs dénotent la cooccurrence de ces termes. Ce graphe peut être perçu comme un graphe de recommandation où l'apparition de deux termes dans un même document signifie qu'ils se recommandent l'un l'autre. Cette observation nous amène naturellement à utiliser un algorithme de marche aléatoire (*random walk* (Cohen, 2010; Page *et al.*, 1999)) attribuant une pertinence globale à chaque sommet du graphe.

Ce critère de cohésion thématique peut avoir de multiples applications. Dans le cadre de lexiques spécialisés construits automatiquement (Baroni et Bernardini, 2004; Groc *et al.*, 2011), ordonner les éléments du lexique par leur score de cohésion peut réduire la charge de validation manuelle ou même limiter la dispersion au fil des itérations. Dans le cadre de la traduction assistée, une valeur de cohésion peut représenter un score de confiance utile pour le traducteur. C'est d'ailleurs par cette dernière application que nous choisissons d'évaluer notre critère dans cet article.

L'article présente tout d'abord brièvement les travaux en *Web as corpus* dont notre approche découle, ainsi que ceux centrés sur les graphes de cooccurences et les algorithmes de marche aléatoire (section 2). Dans une 3ème section, nous présentons le modèle de graphe et l'algorithme de marche aléatoire utilisés pour le calcul du critère de cohésion thématique. Nous évaluons ensuite ce dernier sur une tâche de filtrage de lexiques thématiques traduits automatiquement (section 4). Nous concluons enfin (section 5) en suggérant plusieurs pistes envisagées.

2 Travaux liés

L'utilisation du Web comme source de documents (Kilgarriff et Grefenstette, 2003) est une idée maintenant largement répandue. Pour accéder aux documents Web, deux approches sont couramment mises en œuvre : soumettre un ensemble de requêtes à un moteur de recherche (Baroni et Bernardini, 2004; Ghani *et al.*, 2005) ou parcourir directement le Web (*crawling*) (Baroni et Ueyama, 2006). Le parcours du Web permet une meilleure spécification du besoin mais nécessite un investissement important. De plus, les efforts des moteurs de recherche pour garantir des résultats de qualité doivent être reproduits (filtrage des pages de spam). Au contraire, l'utilisation d'un moteur de recherche grand public comme point d'entrée au Web offre un accès simple et peu coûteux pour la communauté de Traitement Automatique des Langues. Nous adoptons ici cette approche afin de constituer un ensemble de connaissances exogènes sur les lexiques thématiques fournis en entrée.

Dans cet article, nous définissons un critère basé sur les cooccurrences des termes d'une même thématique pour déterminer leur lien avec le thème. Ces travaux partagent donc certaines hypothèses avec les travaux en similarité sémantique et notamment les analyses distributionnelles (Pereira *et al.*, 1993; Baker et McCallum, 1998; Rajman *et al.*, 2000) ou la désambiguïsation sémantique *via* des réseaux de cooccurrences (Dorow et Widdows, 2003; Ferret, 2004). En effet, notre graphe de cooccurrences modélise explicitement les cooccurrences de premier ordre mais l'application d'un algorithme de propagation d'importance de type PageRank permet la prise en compte de cooccurrences d'ordres supérieurs.

Enfin, l'algorithme TextRank (Mihalcea et Tarau, 2004) est étroitement lié à nos travaux. Les auteurs modélisent la cooccurrence des mots dans une fenêtre de taille N sous forme de graphe non-orienté et appliquent un algorithme de marche aléatoire afin de détecter les mots-clés saillants. Nous nous démarquons cependant de ces travaux en au moins deux points : nous considérons les cooccurrences au niveau du document (*snippet* dans nos évaluations) et modélisons ces dernières par un graphe orienté (plus de détails en Section 3).

3 Un critère de cohésion thématique

Étant donné un lexique thématique $\mathscr{L}_T$ composé de N termes, $\mathscr{L}_T = (t_1, t_2, \ldots, t_N)$, nous voulons calculer un vecteur de poids $\mathbf{w}_{\mathscr{L}_T} = (w_1, w_2, \ldots, w_N)$ où chaque poids w_i mesure la pertinence du terme t_i pour la thématique T.

3.1 Recueil de connaissances exogènes

Dans ces travaux, nous adoptons une approche *Web as corpus*, qui nous permet de créer rapidement des corpus spécialisés en nous appuyant sur un moteur de recherche généraliste. Nous proposons dès lors de constituer, pour chaque terme t_i, un corpus C_i correspondant au M meilleurs résultats renvoyés par un moteur de recherche pour la requête `"<t_i>"`.

L'unité d'information que nous considérons dans le cadre de cet article est le *snippet*, le court extrait de page Web renvoyé par le moteur de recherche. En effet, si prendre en compte le document entier permettrait en théorie de bénéficier d'un contexte plus large et plus riche, cela

pose surtout de nombreux problèmes. D'une part, télécharger les documents renvoyés par le moteur de recherche rallonge considérablement le temps de calcul. D'autre part, il est ensuite indispensable d'opérer un nettoyage des pages Web, et en particulier de supprimer les menus, les publicités ou les balises HTML, dans le but de ne conserver que le minimum de contenu non informationnel [1]. L'évaluation finale dépend donc beaucoup de la qualité de ce nettoyage, ce qui la rend plus difficilement interprétable. Enfin, le caractère local des *snippets* peut permettre de réduire le bruit pouvant apparaître dans les pages Web.

Nous avons utilisé le moteur de recherche Bing [2] comme source de *snippets*. Ces derniers sont composés de portions de textes de 155 caractères en moyenne issus du corps des pages Web et contenant les termes de la requête.

3.2 Cohésion thématique et graphe de cooccurrences

Étant donné un lexique thématique $\mathscr{L}_T$, nous proposons une première définition de notre critère comme suit : le poids w_i d'un terme t_i est égal au nombre de termes du lexique (t_i exclu) cooccurrant avec t_i dans le corpus C_i. Plus formellement, le poids w_i d'un terme t_i est défini par :

$$w_i = \sum_{t_j \in \mathscr{L}_T^\dagger} n_{t_j, C_i} \tag{1}$$

où n_{t_j,C_i} est le nombre d'occurrences du terme t_j dans l'ensemble des documents du corpus C_i et $\mathscr{L}_T^\dagger = \mathscr{L}_T \setminus \{t_i\}$, c'est-à-dire l'ensemble des termes du lexique $\mathscr{L}_T$, t_i exclu.

Cette même définition peut être modélisée sous la forme d'un graphe (Figure 1). Soit un graphe orienté $G =< V, E >$, où V est l'ensemble des sommets ($V = \mathscr{L}_T$) et E l'ensemble des arcs. Chaque arc $e(t_i, t_j)$ symbolise l'apparition du terme t_i dans le corpus C_j de t_j. Les arcs sont pondérés en fonction du nombre d'occurrences du terme t_i dans C_j. Notre approche *Web as corpus* nous différencie des précédents travaux (Mihalcea et Tarau, 2004) visant à modéliser les cooccurrences sous forme de graphe non-orienté : en effet, pour deux termes t_i et t_j et leurs corpus associés C_i et C_j, l'apparition du terme t_i dans le corpus C_j constitue un indice du "vote" de t_i pour t_j. Cependant, cette relation n'est pas symétrique puisque les corpus C_i et C_j sont distincts. En conséquence, nous optons pour un modèle de graphe orienté.

Le poids d'un terme tel que défini par l'équation 1 est alors équivalent au degré entrant de ce terme dans le graphe, c'est-à-dire la somme des poids des arcs entrants.

Cette nouvelle modélisation graphique nous amène à intégrer les poids du voisinage entrant d'un sommet dans le calcul du poids de celui-ci. Nous rectifions alors la première définition de notre critère et proposons la définition suivante : le poids w_i d'un terme t_i est égal à la somme des poids des termes du lexique cooccurrant avec t_i dans le corpus C_i. De plus, nous normalisons cette somme afin que le poids d'un terme soit réparti entre tous les termes auxquels il est lié.

1. Ce problème est d'ailleurs un thème de recherche à part entière fédéré par la campagne d'évaluation CLEANEVAL (`http://cleaneval.sigwac.org.uk`).
2. `http://www.bing.com`

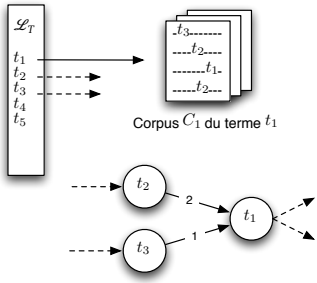

FIGURE 1 – Modélisation des cooccurrences sous forme de graphe orienté

Formellement, nous définissons le poids w_i d'un terme t_i comme :

$$w_i = \frac{\sum_{t_j \in \mathscr{L}_T^\dagger} n_{t_j,C_i} \cdot w_j}{\sum_{t_j \in \mathscr{L}_T^\dagger} n_{t_j,C_i}} \tag{2}$$

Cette nouvelle définition est "récursive" dans le sens où la pertinence d'un terme du lexique est définie en fonction de la pertinence des autres termes du lexique apparaissant dans son corpus. Il est ainsi possible de voir la pertinence d'un terme t_i défini en fonction de la pertinence d'un terme t_j, elle-même défini en fonction la pertinence de t_i. D'un point de vue graphique, ce phénomène se traduit alors simplement par un cycle dans le graphe de cooccurrences.

L'équation 2 est en fait proche de l'algorithme de marche aléatoire PageRank (Page *et al.*, 1999) et peut être résolue par un algorithme itératif sous certaines conditions. Cette version naïve de l'algorithme pose cependant deux problèmes :

1. L'algorithme ne gère pas correctement les sommets sans arcs sortant (appelés "dangling nodes" ou "rank sink" dans la littérature) : il n'est pas souhaitable qu'un terme ne renvoyant que des documents extérieurs à la thématique obtienne un poids important. Par exemple, si un terme de notre lexique $\mathscr{L}_T$ est un mot outil ("et"), il possèdera de nombreux liens entrant mais potentiellement aucun lien sortant : il accumulera itérativement un poids important.
2. La convergence vers une unique solution n'est pas garantie pour notre graphe (Langville et Meyer, 2005; Farahat *et al.*, 2006).

Pour résoudre le premier problème, nous ajoutons un lien des sommets sans arcs sortant vers un sommet virtuel et un lien de ce sommet virtuel vers tous les sommets du graphe. Le poids des sommets sans arcs sortant est ainsi redistribué à tous les sommets du graphe. Concernant le second problème, nous appliquons la solution de Page et Brin (Page *et al.*, 1999) et ajoutons une probabilité de téléportation uniforme à chaque itération de l'algorithme ce qui garantit la forte connectivité du graphe et la convergence vers une solution unique (équation 3).

$$w_{i,n+1} = \frac{(1-\alpha)}{N} + \alpha \cdot \frac{\sum_{t_j \in \mathscr{L}_T^{\dagger}} n_{t_j,C_i} \cdot w_{j,n}}{\sum_{t_j \in \mathscr{L}_T^{\dagger}} n_{t_j,C_i}} \tag{3}$$

où N est le nombre de sommets du graphe (c'est à dire le nombre de termes du lexique) et α est un facteur d'amortissement traditionnellement fixé à 0.85 (Page *et al.*, 1999; Mihalcea et Tarau, 2004). Nous utilisons également cette valeur de α pour nos expériences.

L'algorithme 1 récapitule l'intégralité du calcul du critère de cohésion thématique.

Algorithme 1 Critère de cohésion thématique

Entrées : $\mathscr{L}_T$: termes $t_i, i \in [1, N]$
M : nombre de documents téléchargés par requête
α : facteur d'amortissement

// Téléchargement du corpus
Pour tout terme $t_i \in \mathscr{L}_T$ **faire**
  Soumettre t_i à un moteur de recherche
  Télécharger M documents comme corpus C_i
Fin Pour

// Initialisation
Pour tout terme $t_i \in \mathscr{L}_T$ **faire**
  $w_{i,1} = 1/N$
Fin Pour

// Procédure itérative de calcul des poids
$n = 1$
Tant que (non convergence) **faire**
  Pour tout terme $t_i \in \mathscr{L}_T$ **faire**
    $w_{i,n+1} = \frac{(1-\alpha)}{N} + \alpha \cdot \frac{\sum_{t_j \in \mathscr{L}_T^{\dagger}} n_{t_j,C_i} \cdot w_{j,n}}{\sum_{t_j \in \mathscr{L}_T^{\dagger}} n_{t_j,C_i}}$
  Fin Pour
  Normalisation des poids : $\sum_i w_{i,n+1} = 1$
  $n = n + 1$
Fin Tant que
Retourner w_n

Pour améliorer la correspondance entre les lexiques et les documents issus du Web, une série de normalisations supplémentaires est appliquée : conversion des termes en minuscules, racinisation (*stemming*) et normalisation des caractères unicode (accents, ...).

Notons que le critère proposé traite les phénomènes d'ambiguïtés graphiques de la langue (homographie) de la façon souhaitée. Par exemple, si le terme "jaguar" est soumis à un moteur de recherche actuel, il est fort probable que ce dernier renvoie des résultats diversifiés à propos de l'animal mais également de la marque de voiture, de la console de jeu ou du système d'exploitation MacOS. Le nombre de cooccurrences avec les termes du lexique sera donc plus limité, conduisant à un score plus faible, soulignant ainsi qu'un terme ambigüe contribue moins à la cohésion thématique.

4 Évaluation

4.1 Tâche

Dans cet article, nous évaluons l'apport de notre critère pour l'aide à la création de lexiques thématiques bilingues à partir de lexiques monolingues. Comme mentionné dans l'introduction, nous envisageons de nombreuses applications pour le critère proposé dont notamment le *boots-trapping* de lexiques thématiques monolingues. Cependant, la tâche de création de lexiques thématiques bilingues, claire et facilement reproductible, fournit une évaluation objective de notre critère de cohésion.

Partant de lexiques thématiques monolingues, une approche commune pour la création de lexiques thématiques bilingues est d'utiliser un outil de traduction automatique en ligne tel que Google Translate [3]. Cependant, ces outils ne permettent pas d'intégrer une notion de *contexte thématique* dans le processus de traduction simplement. Ainsi, le terme simple "avocat" non intégré à une phrase, par exemple, sera traduit par ces outils "avocado" ou "lawyer" en anglais, indifféremment du fait qu'il appartient à un lexique juridique ou culinaire. Une validation manuelle laborieuse est donc nécessaire pour supprimer les traductions erronées.

Nous proposons d'appliquer notre critère de cohésion thématique aux lexiques traduits, attribuant ainsi à chaque traduction un score de confiance. Le tri des lexiques en fonction de ce score permet alors de réduire le temps nécessaire à leur validation.

4.2 Données

Nous avons utilisé trois lexiques bilingues français/anglais spécialisés sur trois thèmes différents :
– Astronomie (*The Astronomy Thesaurus* [4]) ;
– Médical (*Unified Medical Language System* - UMLS [5]) ;
– Statistiques (*International Statistical Institute* [6]).
Un exemple de termes issus de chaque lexique bilingue est donné Table 2.

3. http://translate.google.com
4. http://msowww.anu.edu.au/library/thesaurus/
5. http://www.nlm.nih.gov/research/umls/
6. http://isi.cbs.nl/glossary/

Astronomie		Statistiques	
Anglais	*Français*	*Anglais*	*Français*
afterglow	rémanence	Birnbaum's inequality	inégalité de Birnbaum
celestial coordinates	coordonnee celeste	geometric mean	moyenne géométrique
asteroids	asteroide	K-test	test K de Mann
dwarf stars	etoile naine	invariant	invariant
bow shocks	onde de choc en form d'arc	cross spectrum	spectre croisé

Médical	
Anglais	*Français*
wandering spleen	rate flottante
dimethoxyphenylethylamine	diméthoxyphényléthylamine
wolman disease	maladie de wolman
antimalarials	antipaludiques
optical illusions	illusions optiques

TABLE 2 – Extrait des lexiques thématiques bilingues utilisés pour l'évaluation

Une série de traitements a été appliquée à chaque lexique dans le but d'en améliorer la qualité ou l'utilisation pour notre évaluation. Ainsi, nous avons supprimé les termes apparaissant entre crochets ou parenthèses dans les lexiques Astronomie et Statistiques. Le lexique Médical présentant des termes trop ambigus pour être nettoyés automatiquement (par exemple *3-pyridinecarboxylic acid, 1,4-dihydro-2,6-dimethyl-5-nitro-4-(2-(trifluoromethyl)phenyl)-, methyl ester*), nous avons simplement supprimé les termes contenant une parenthèse ou une virgule.

Nous avons ensuite traité le cas des traductions multiples de la manière suivante : lorsqu'un terme de la langue source possédait plusieurs traductions dans la langue cible, nous n'avons conservé que le terme le plus proche (au sens de la distance de Damerau-Levenshtein (Damerau, 1964; Levenshtein, 1966)) de la traduction automatique [7]. Ainsi dans l'exemple "Afterglow" ⇒ "Postluminescence ou Remanence" issu du lexique "Astronomie", nous n'avons conservé que le terme "Remanence" car il est le plus proche de la traduction automatique trouvée : "rémanence".

Le lexique UMLS comprenant plus de 19 000 termes, nous avons choisi de ne travailler que sur un échantillon de ce dernier. Nous avons donc tiré aléatoirement deux séries de 2 000 termes que nous désignons comme lexiques Médical-1 et Médical-2.

Nous obtenons enfin les lexiques suivants :
– Astronomie (2 940 termes) ;
– Statistiques (2 752 termes) ;
– Médical-1 (2 000 termes) ;
– Médical-2 (2 000 termes).

4.3 Méthode

Nous traduisons chaque lexique thématique d'une langue source vers une langue cible (par exemple Astronomie fr → en ou Astronomie en → fr) à l'aide du moteur de traduction Google

7. voir section 4.3 pour la méthode de traduction automatique

Translate. Le lexique résultant est alors pondéré avec le critère de cohésion thématique puis ordonné et comparé avec la référence.

Le choix de Google Translate est justifié par le fait que ce moteur propose une très large couverture, ce qui en fait un candidat idéal pour traiter nos lexiques spécialisés. De plus, les récentes évaluations du NIST ont montré que l'outil de Google propose des performances état de l'art quant à la qualité des traductions produites (NIST, 2005, 2008).

La comparaison entre les termes traduits et les termes des lexiques originaux est réalisée à l'aide d'une mesure ad hoc incluant la suppression des déterminants en début de terme ("le bleu de bromothymol" ⇒ "bleu de bromothymol") et une distance d'édition de Damerau-Levenshtein (Damerau, 1964; Levenshtein, 1966). Nous avons considéré un terme comme valide s'il est au plus à une distance d'édition de 1 du terme de référence, autorisant ainsi une légère marge d'erreur due au moteur de traduction ou à la référence (singuliers transformés en pluriels, espace remplacé par un tiret, . . .).

La mesure de précision moyenne non-interpolée (*uninterpolated average precision* - UAP (Manning et Schütze, 1999)) est employée pour évaluer la validité de l'ordre des termes traduits.

4.4 Évaluation du critère

Nous évaluons notre critère *Cohésion-RW* comparativement à une baseline *Hasard*, obtenue par le simple tri aléatoire de la liste de traduction, et à la première version du critère *Cohésion-DEG* (équation 1). La baseline *Hasard* est obtenue en calculant une macro-moyenne sur dix tris aléatoires successifs. Nous fixons le nombre de *snippets* téléchargés pour chaque requête (la valeur M de l'algorithme 1) à 200 documents. Les résultats sont présentés à la Table 3.

Thème		Hasard	Cohésion-DEG	Cohésion-RW
Astronomie	en → fr	0.429	0.494	**0.512**
	fr → en	0.553	0.664	**0.678**
Statistiques	en → fr	0.382	0.663	**0.711**
	fr → en	0.488	0.667	**0.705**
Médical-1	en → fr	0.530	0.683	**0.735**
	fr → en	0.620	0.707	**0.718**
Médical-2	en → fr	0.522	0.662	**0.699**
	fr → en	0.638	0.739	**0.750**

TABLE 3 – Précision moyenne non-interpolée (UAP) pour le classement des termes des lexiques traduits.

Nous constatons que l'algorithme de marche aléatoire fournit les meilleurs résultats avec gain sur la baseline *Hasard* allant de 15,8 % (Médical-1 fr → en) à 86,1 % (Statistiques en → fr). La baseline fournit une idée de la qualité des traductions produites par le moteur de traduction. Ainsi, il semble que les lexiques Médical-1 et Médical-2 soient les mieux traduits. Au contraire le lexique Statistiques semble être le plus difficile à traduire. Le coefficient de corrélation de Pearson entre la précision du Hasard et le gain obtenu vaut -0,61 ce qui semble signifier qu'ils sont fortement corrélés négativement : plus la traduction est de bonne qualité et plus le gain est

Thème		50	100	150	200
Astronomie	en → fr	0.500	0.507	0.507	**0.512**
	fr → en	0.672	0.676	**0.680**	0.678
Statistiques	en → fr	0.666	0.695	0.704	**0.711**
	fr → en	0.678	0.691	0.702	**0.705**
Médical-1	en → fr	0.710	0.719	0.726	**0.735**
	fr → en	0.677	0.693	0.706	**0.718**
Médical-2	en → fr	0.672	0.678	0.687	**0.699**
	fr → en	0.702	0.723	0.735	**0.750**

TABLE 4 – Précision moyenne non-interpolée (UAP) pour le classement des termes des lexiques traduits avec le critère Cohésion-RW pour différentes valeurs de NB_DOCS.

faible. Cependant ce résultat n'est pas statistiquement significatif (la *p-value* vaut 0,106).

Nous évaluons ensuite l'influence du nombre de *snippets* téléchargés par requête sur les résultats du critère proposé (Table 4). Nous constatons que la précision moyenne augmente avec le nombre de documents téléchargés. Cela est probablement dû au fait que la qualité du graphe de cooccurrences augmente avec le nombre de documents et que la précision moyenne en est directement impactée.

4.5 Stabilité du critère

Le poids d'un terme t_i est défini en fonction des cooccurrences du terme t_i avec les autres termes du lexique (équation 3). Il semble donc légitime de s'interroger quant à la stabilité des poids des termes en fonction de la taille du lexique. La somme des poids étant égale à 1, le poids absolu de chaque terme est donc lié à la taille du graphe, il va diminuer avec l'augmentation du nombre total de termes. La question est donc de savoir ce qu'il en est du poids relatif, c'est-à-dire du rang des termes en fonction de la taille des lexiques.

Pour évaluer l'évolution des rangs des termes, nous avons de nouveau utilisé les lexiques traduits. Pour chaque lexique, nous avons sélectionné aléatoirement 20 termes, puis avons augmenté itérativement la taille du lexique de 20 termes. Chaque lexique (de 20, 40, 60, ... termes) a ensuite été ordonné par le critère de cohésion thématique. Enfin, le coefficient de corrélation de Spearman a été employé pour mesurer l'évolution des rangs des termes entre lexiques successifs (20-40, 40-60, ...).

Afin de réduire l'influence du hasard, nous avons répété la procédure décrite précédemment 10 fois et avons calculé une macro-moyenne des coefficients de corrélation. Dans un soucis de clarté, nous ne présentons que les résultats sur les lexiques anglais (Figure 2). Toutefois, les résultats obtenus sur les lexiques français sont équivalents.

Nous constatons que le coefficient de corrélation augmente pour tous les lexiques au fur et à mesure que la taille des lexiques augmente. Déjà forte avec une corrélation de plus de 0,70, le coefficient de corrélation s'approche du maximum à partir de 300 termes. Autrement dit, passée cette limite, l'ajout de nouveaux termes ne modifie presque pas l'ordre déjà établi entre les autres termes, ce qui nous permet de conclure que la mesure est stable à partir de ce seuil.

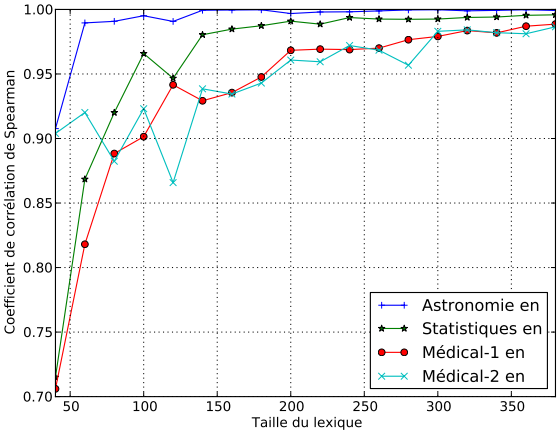

FIGURE 2 – Étude de l'évolution des rangs des termes en fonction de la taille des lexiques

5 Conclusion

Dans cet article, nous avons présenté un nouveau critère de cohésion thématique fondé sur un graphe de cooccurrences et un algorithme de marche aléatoire.

Nous avons évalué ce critère par une tâche d'aide à la création de lexiques bilingues car il s'agit d'une tâche claire, facilement reproductible et évaluable de façon objective. Cependant, comme nous l'avons indiqué, les applications possibles sont diverses, que ce soit pour réduire la charge de validation manuelle ou pour mieux sélectionner les termes automatiquement pour de la recherche d'information, de la collecte de corpus ou la mise en œuvre de techniques de *bootstrapping*. Les résultats obtenus sont encourageants et montrent la pertinence de notre approche.

Nous prévoyons d'analyser plus en détail le comportement du critère proposé en évaluant, par exemple, sa robustesse à la présence de termes non-pertinents dans les lexiques thématiques. Nous comptons également évaluer l'apport de quelques annotations manuelles en intégrant ces annotations dans l'algorithme de marche aléatoire sous forme d'un vecteur de personnalisation (Haveliwala, 2003).

Remerciements

Nous voudrions remercier Pierre Zweigenbaum de nous avoir fourni les lexiques nécessaires à l'évaluation de notre méthode et l'International Statistical Institute de nous avoir autorisé à utiliser le glossaire de termes statistiques multilingue. Nous remercions également Javier Couto pour ses conseils avisés sur la première version de ce manuscrit ainsi que les relecteurs anonymes pour leurs remarques et conseils. Ce travail s'inscrit dans le cadre des projets METRICC (ANR-08-CORD-013) et TTC (FP7/2007-2013 GA n°248005).

Références

BAKER, L. et MCCALLUM, A. (1998). Distributional clustering of words for text classification. *In Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 96–103.

BARONI, M. et BERNARDINI, S. (2004). BootCaT : Bootstrapping Corpora and Terms from the Web. *In Proceedings of the LREC 2004 conference*, pages 1313–1316.

BARONI, M. et UEYAMA, M. (2006). Building general-and special-purpose corpora by web crawling. *In Proceedings of the 13th NIJL international symposium, language corpora : Their compilation and application*, pages 31–40.

COHEN, W. W. (2010). *Graph Walks and Graphical Models*. Carnegie Mellon University, School of Computer Science, Machine Learning Dept.

DAMERAU, F. J. (1964). A technique for computer detection and correction of spelling errors. *Communications of the ACM*, 7(3):171–176.

DOROW, B. et WIDDOWS, D. (2003). Discovering corpus-specific word senses. *In Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics-Volume 2*, pages 79–82.

FARAHAT, A., LOFARO, T., MILLER, J., RAE, G. et WARD, L. (2006). Authority rankings from hits, pagerank, and salsa : Existence, uniqueness, and effect of initialization. *SIAM Journal on Scientific Computing*, 27(4):1181–1201.

FERRET, O. (2004). Discovering word senses from a network of lexical cooccurrences. *In Proceedings of the 20th international conference on Computational Linguistics*, pages 1326–1332.

GHANI, R., JONES, R. et MLADENIC, D. (2005). Building Minority Language Corpora by Learning to Generate Web Search Queries. *Knowl. Inf. Syst.*, 7(1):56–83.

GROC, C. d., TANNIER, X. et COUTO, J. (2011). GrawlTCQ : Terminology and Corpora Building by Ranking Simultaneously Terms , Queries and Documents using Graph Random Walks. *In Proceedings of the TextGraphs-6 Workshop, Association for Computational Linguistics*, pages 37–41.

HAVELIWALA, T. (2003). Topic-sensitive pagerank : A context-sensitive ranking algorithm for web search. *Knowledge and Data Engineering, IEEE Transactions on*, 15(4):784–796.

KILGARRIFF, A. et GREFENSTETTE, G. (2003). Introduction to the Special Issue on the Web as Corpus. *Computational Linguistics*, 29(3):333–347.

LANGVILLE, A. et MEYER, C. (2005). A survey of eigenvector methods for web information retrieval. *SIAM review*, pages 135–161.

LEVENSHTEIN, V. (1966). Binary Codes Capable of Correcting Deletions, Insertions and Reversals. *Soviet Physics Doklady*, 10:707–710.

MANNING, C. et SCHÜTZE, H. (1999). *Foundations of statistical natural language processing*. MIT Press.

MIHALCEA, R. et TARAU, P. (2004). TextRank bringing order into text. *In Proceedings of the Conference on Empirical Methods on Natural Language Processing*, pages 404–411.

NIST (2005). Nist 2005 machine translation evaluation official results. `http://www.itl.nist.gov/iad/mig/tests/mt/2005/doc/mt05eval_official_results_release_20050801_v3.html`. [consulté le 23/01/2012].

NIST (2008). Nist 2008 open machine translation evaluation (mt08) - official evaluation results. `http://www.itl.nist.gov/iad/mig/tests/mt/2008/doc/mt08_official_results_v0.html`. [consulté le 23/01/2012].

PAGE, L., BRIN, S., MOTWANI, R. et WINOGRAD, T. (1999). The PageRank Citation Ranking : Bringing Order to the Web. Rapport technique, Stanford InfoLab.

PEREIRA, F., TISHBY, N. et LEE, L. (1993). Distributional clustering of english words. *In Proceedings of the 31st annual meeting on Association for Computational Linguistics*, pages 183–190.

RAJMAN, M., BESANÇON, R. et CHAPPELIER, J. (2000). Le modèle dsir : Une approche à base de sémantique distributionnelle pour la recherche documentaire. *Traitement automatique des langues*, 41(2):549–578.

WAN, X. (2009). Co-training for cross-lingual sentiment classification. *In Proceedings of the Joint Conference of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the AFNLP : Volume 1*, pages 235–243.