Modélisation à base de graphe pour l'indexation en domaines de spécialité

A graph-based ranking approach for indexing in specialised domains

Adrien Bougouin, Florian Boudin, Béatrice Daille

LS2N- Univ. Nantes, 2 rue de la Houssinière, 44322 Nantes Cedex 3, France
prenom.nom@univ-nantes.fr

RÉSUMÉ. Cet article présente la participation de l'équipe TALN du LINA au défi fouille de textes (DEFT) 2016. Pour la tâche d'indexation de documents de domaines de spécialité par l'intermédiaire de leurs mots-clés, nous avons proposé une méthode à base de graphe, TopicCoRank, dans la lignée des approches à base de graphes proposées en recherche d'information. TopicCoRank modélise les informations présentes dans le document et la connaissance du domaine pour réaliser une indexation plus exhaustive et respectueuse du vocabulaire du domaine. Notre système s'est classé à la troisième place quel que soit le domaine de spécialité.

ABSTRACT. This article presents the participation of the TALN group at LINA to the défi fouille de textes (DEFT) 2016. Developed specifically for automatic keyphrase annotation, we propose a new method, TopicCoRank, extracting the most important phrases from a document and providing key-phrases that do not occur in the document. Our system ranked third out of a total of five systems.

MOTS-CLÉS. DEFT 2016, extraction de mots-clés, assignation de mots-clés, méthode à base de graphe, domaine de spécialité.

KEYWORDS. DEFT 2016, keyphrase extraction, keyphrase assignment, graph-based method, specific domain.

1 Introduction

L'indexation automatique consiste à identifier un ensemble de mots-clés (e.g. mots, termes, entités nommées, expressions multi-mots) qui décrit le contenu d'un document. Les mots-clés peuvent ensuite être utilisés, entre autres, pour faciliter la recherche d'information ou la navigation dans les collections de documents. L'édition 2016 du défi fouille de textes (DEFT) porte sur l'extraction automatique de mots-clés à partir d'articles scientifiques en français. La tâche consiste à retrouver, à partir du contenu des notices d'articles scientifiques, les mots-clés qui ont été attribués par des indexeurs professionnels. Les documents sont issus de quatre domaines de spécialité : la linguistique, les sciences de l'information, l'archéologie et la chimie.

Il existe de nombreuses méthodes pour l'extraction automatique de mots-clés. Certaines se fondent uniquement sur des statistiques et d'autres les combinent avec des représentations plus complexes du document comme des groupes sémantiques ou des graphes de cooccurrences de mots. Nous avons utilisé une méthode à base de graphe, TopicRank, qui minimise la redondance et l'imprécision dans l'extraction des mots-clés. Son adaptation, TopicCoRank, tire profit des données déjà indexées et réalise une indexation libre et contrôlée.

Nous commençons par présenter l'état de l'art des méthodes à base de graphe (cf. section 2), la méthode à base de graphe TopicRank qui effectue une extraction non-supervisée de mots-clés et ses résultats obtenus par rapport aux méthodes non supervisées à base de graphe état de l'art sur les données de DEFT 2016 (cf. section 3), puis la méthode à base de graphe supervisée TopicCoRank qui adapte TopicRank et exploite les notices déjà indexées du corpus d'entraînement (cf. section 4). Enfin, nous indiquons les méthodes soumises à DEFT 2016 (cf. section 5), rappelons les résultats de TopicCoRank comparés à ceux des autres participants (cf. section 6) et faisons une analyse des principales erreurs (cf. section 7) avant de conclure.

Méthode	DUC *(Wan & Xiao, 2008)*	Inspec *(Hulth, 2003)*	NUS *(Nguyen & Kan, 2007)*	ICSI *(Adam* et al.*, 2003)*
TF-IDF*	27,0	36,3	**6,6**	**12,1**
TextRank*	9,7	33,0	3,2	2,7
SingleRank*	25,6	35,3	3,8	4,4
ExpandRank*	26,9	35,3	3,8	4,3
TopicalPageRank	31,2	—	—	—
WordTopic-MultiRank	**34,0**	**48,2**	—	—

Tableau 2.1. *Comparaison des méthodes d'extraction automatique de mots-clés de la littérature, lorsque dix mots-clés sont extraits. Les performances sont exprimées en mot de F-mesure.* DUC *est une collection d'articles journalistiques, Inspec est une collection de résumés d'articles scientifiques,* NUS *est une collection d'articles scientifiques et* ICSI *est une collections de transcriptions textuelles de réunions.* * *indique que les résultats donnés par Hasan & Ng (2010).*

2 Méthodes à base de graphe

Les approches à base de graphe sont actuellement très populaires et utilisées dans de nombreuses applications du TAL (Kozareva *et al.*, 2013), les graphes ayant l'avantage de proposer une modélisation simple et intuitive du document.

Mihalcea & Tarau (2004) proposent TextRank, une méthode d'ordonnancement d'unités textuelles à partir d'un graphe pour le résumé automatique et l'extraction de mots-clés. Pour l'extraction de mots-clés, les nœuds du graphe sont les mots du document et les arêtes qui les connectent représentent leurs relations d'adjacence calculées dans une fenêtre de deux mots dans le document. Un score d'importance (initialisé à un) est calculé pour chaque mot à partir de l'algorithme itératif PageRank (Brin & Page, 1998). PageRank est un algorithme de marche aléatoire (*random walk*) : un marcheur aléatoire parcoure le graphe de mot en mot. Le résultat du parcours du marcheur permet de déduire l'importance de chaque mot d'après le principe de la recommandation (du vote) : un mot est d'autant plus important s'il apparaît avec un grand nombre de mots (parce qu'il est beaucoup visité par le marcheur) et si les mots avec lesquels il apparaît sont eux aussi importants (parce qu'il a plus de chance d'être visité par le marcheur). Les mots les plus importants sont marqués dans le document et les plus longues séquences de mots importants sont extraites en tant que mots-clés.

Soit le graphe de cooccurrences de mots non orienté $G = (N, A)$, où les nœuds N représentent les mots du documents, et où les arêtes A les connectent lorsqu'ils apparaissent ensemble dans le document. L'importance de chaque mot n_i est obtenue itérativement selon la formule TextRank suivante :

$$S(n_i) = (1 - \lambda) + \lambda \times \sum_{n_j \in A(n_i)} \frac{S(n_j)}{|A(n_j)|} \quad [1]$$

où $A(n_i)$ est l'ensemble des nœuds connectés au nœud n_i et où λ est un facteur d'atténuation. Nombre réel défini entre 0 et 1, ce dernier peut être considéré comme la probabilité pour que le nœud n_i soit important d'après le principe de la recommandation. Brin & Page (1998) suggèrent 0,85 comme valeur par défaut de λ. Selon eux, cette valeur est un bon compromis entre la précision des résultats et la vitesse de convergence de l'algorithme.

Bien qu'intéressant, de par son intuitivité, Hasan & Ng (2010) ont montré que TextRank est moins performant que TF-IDF (Salton *et al.*, 1975) (cf. le tableau 2.1).

Wan & Xiao (2008) modifient TextRank et proposent SingleRank. Dans un premier temps, leur méthode augmente la précision de l'ordonnancement en utilisant une fenêtre de collecte des cooccurrences élargie empiriquement à dix mots et en pondérant les arêtes par le nombre de cooccurrences des deux mots qu'elles connectent. La pondération, notée poids(n_j, n_i), sert à ajuster l'importance du mot n_i acquise à partir de sa recommandation par le mot n_j (équation 2). Dans un second temps, les mots-clés ne sont plus générés à partir des séquences de mots-clés dans le document, mais ordonnés à partir de la somme du score d'importance des mots qui les composent. Comparé à TextRank, dans les expériences de Hasan & Ng (2010) réalisées avec quatre collections de données différentes,

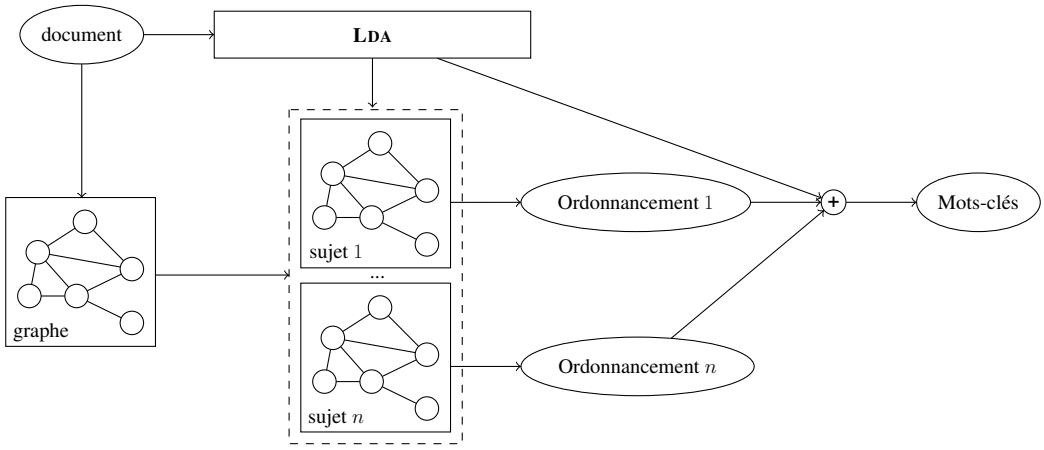


FIGURE 2.1. *Illustration du fonctionnement de TopicalPageRank (Liu et al., 2010).*

SingleRank donne de meilleurs résultats (cf. le tableau 2.1). Ils restent cependant plus faibles que ceux de TF-IDF.

$$S(n_i) = (1-\lambda) + \lambda \times \sum_{n_j \in A(n_i)} \frac{\text{poids}(n_j, n_i) \times S(n_j)}{\sum_{n_k \in A(n_j)} \text{poids}(n_j, n_k)} \quad [2]$$

Toujours dans le but d'améliorer l'efficacité de l'ordonnancement proposé par Mihalcea & Tarau (2004), Wan & Xiao (2008) proposent ExpandRank. ExpandRank étend SingleRank en utilisant des documents similaires au document analysé d'après la mesure de similarité vectorielle cosinus. Faisant l'hypothèse que ces documents similaires fournissent des informations supplémentaires relatives aux mots du document et aux relations qu'ils entretiennent, ExpandRank utilise les relations de cooccurrences observées dans les documents similaires pour ajouter et renforcer des arêtes dans le graphe. Dans leurs expériences réalisée avec une collection de 308 articles journalistiques, Wan & Xiao (2008) obtiennent des résultats au-delà de ceux de SingleRank. Ces résultats n'ont cependant jamais pu être reproduit et les expériences de Hasan & Ng (2010) ne montrent globalement pas d'amélioration vis-à-vis des resultats de SingleRank (cf. le tableau 2.1).

Liu *et al.* (2010) tentent aussi d'améliorer SingleRank. Ils proposent TopicalPageRank (TPR), une méthode qui cherche cette fois-ci à augmenter la couverture du document par les mots-clés extraits. Pour ce faire, ils détectent les sujets du document et ordonnent les mots en fonction de chaque sujet (cf. figure 2.1). À l'aide du modèle d'analyse sémantique latente LDA (Blei *et al.*, 2003), ils ajustent chaque ordonnancement avec la probabilité conditionnelle d'un sujet donné sachant chaque mot (équation 3), puis donnent plus d'importance aux candidats dont les mots ont la plus forte importance (le meilleur rang) selon les sujets les plus probables dans le document (équation 4).

$$S(n_i, \textit{sujet}) = (1-\lambda) \times p(\textit{sujet}|n_i) + \lambda \times \sum_{n_j \in A(n_i)} \frac{\text{poids}(n_j, n_i) \times S(n_j)}{\sum_{n_k \in A(n_j)} \text{poids}(n_j, n_k)} \quad [3]$$

$$\text{TPR}(\textit{candidat}) = \sum_{\textit{sujet}} \left[p(\textit{sujet}|\textit{document}) \times \sum_{n \in \textit{candidat}} \text{rang}_{\textit{sujet}}(n) \right] \quad [4]$$

Contrairement aux précédentes méthodes à base de graphe que nous avons présentées, les résultats de TopicalPageRank surpassent ceux de TF-IDF (cf. tableau 2.1).

Dans la continuité du travail de Liu *et al.* (2010), Zhang *et al.* (2013) proposent WordTopic-MultiRank. WordTopic-MultiRank ajoute les sujets de LDA aux nœuds du graphe de cooccurrences et effectue un seul ordonnancement, qui tient compte de tous les sujets en même temps. Cet ordonnancement est réalisé conjointement entre les mots et les sujets, de sorte que :

— un sujet est d'autant plus important s'il est connecté à un grand nombre de mots importants ;
— un mot est d'autant plus important s'il apparaît dans le même contexte avec un grand nombre de mots importants et s'il est connecté à un grand nombre de sujets importants.

Comme pour SingleRank et TopicalPageRank, les mots-clés candidats sont ensuite ordonnés d'après le score d'importance des mots qu'ils contiennent.

L'ordonnancement conjoint (*co-ranking*) à partir de modèles à base de graphe est une technique qui commence à susciter de l'intérêt en TAL (Wan, 2011; Yan *et al.*, 2012; Liu *et al.*, 2014). Zhang *et al.* (2013) sont les premiers à l'appliquer à l'extraction de mots-clés. Cette approche est intéressante, car elle tient compte à la fois du contexte local du mot (le document) et de son contexte global (les sujets de la collection de données analysée par LDA).

Comparés aux résultats de TopicalPageRank, Zhang *et al.* (2013) montrent que l'ordonnancement conjoint des mots et des sujets est légèrement plus performant que la combinaison de multiples ordonnancements influencés par chaque sujet (cf. tableau 2.1).

L'approche que nous proposons, TopicCoRank, réalise aussi un ordonnancement conjoint entre les mots-clés extraits des notices et les mots-clés présents dans les notices déjà indexées. TopicCoRank s'appuie sur une méthode à base de graphe, TopicRank, qui extrait les mots-clés d'un document et les regroupe en sujet. Nous présentons maintenant TopicRank, puis ses résultats ainsi que ceux des méthodes non supervisées état de l'art TextRank et SingleRank obtenus sur les données DEFT 2016.

3 TopicRank

TopicRank (Bougouin & Boudin, 2014) est une méthode à base de graphe non-supervisée pour l'extraction de mots-clés qui identifie les sujets du document. Un sujet représente un concept véhiculé par un ou plusieurs mots-clés candidats. TopicRank sélectionne en premier les mots-clés candidats au sein du document à analyser, les regroupe pour constituer des sujets, projette ces sujets dans un graphe puis les ordonne en fonction de leur importance en adoptant une marche aléatoire s'inspirant de TextRank (cf. Section 2).

Afin de proposer une méthode générique n'utilisant pas de données supplémentaires, nous appliquons un groupement « naïf » des candidats calculé par similarité lexicale.

Deux candidats c_1 et c_2 sont considérés comme des ensembles de mots (sacs de mots) et leur degré de similarité est calculé à l'aide de la mesure de Jaccard (équation 6), de sorte qu'ils soient très similaires s'ils partagent un grand nombre de mots. Contrairement à d'autres méthodes d'extraction de mots-clés comme TF-IDF, TopicRank est capable de réduire considérablement la redondance des mots-clés extraits (**?**).

TopicRank repose sur cinq grandes étapes :

1. **Sélection des mots-clés candidats.** Suivant les travaux précédents (Wan & Xiao, 2008; Hasan & Ng, 2010), TopicRank sélectionne les plus longues séquences de noms et d'adjectifs en tant que mots-clés candidats :

$$\text{mots\_cles\_candidat} = (NOM|ADJ)+ \quad [5]$$

2. **Groupement en sujets.** TopicRank groupe les mots-clés candidats similaires en sujets. Deux candidats c_i et c_j sont jugés similaires lorsqu'ils partagent au moins un quart de leurs mots, racinisés d'après la méthode de Porter (1980) :

$$\text{sim}(c_i, c_j) = \frac{|\text{Porter}(c_i) \cap \text{Porter}(c_j)|}{|\text{Porter}(c_i) \cup \text{Porter}(c_j)|} \quad [6]$$

$$\forall c_i, c_j \in CANDIDATS, c_j \in \text{sujet}(c_i) \Rightarrow \text{sim}(c_i, c_j) \geq \frac{1}{4} \quad [7]$$

   Le groupement est réalisé avec un algorithme de groupement hiérarchique agglomératif.

3. **Construction du graphe des sujets.** TopicRank représente le document par un graphe complet $G = (N, A \subseteq N \times N)$ où les nœuds N sont les sujets. Chaque sujet $n \in N$ est connecté aux autres

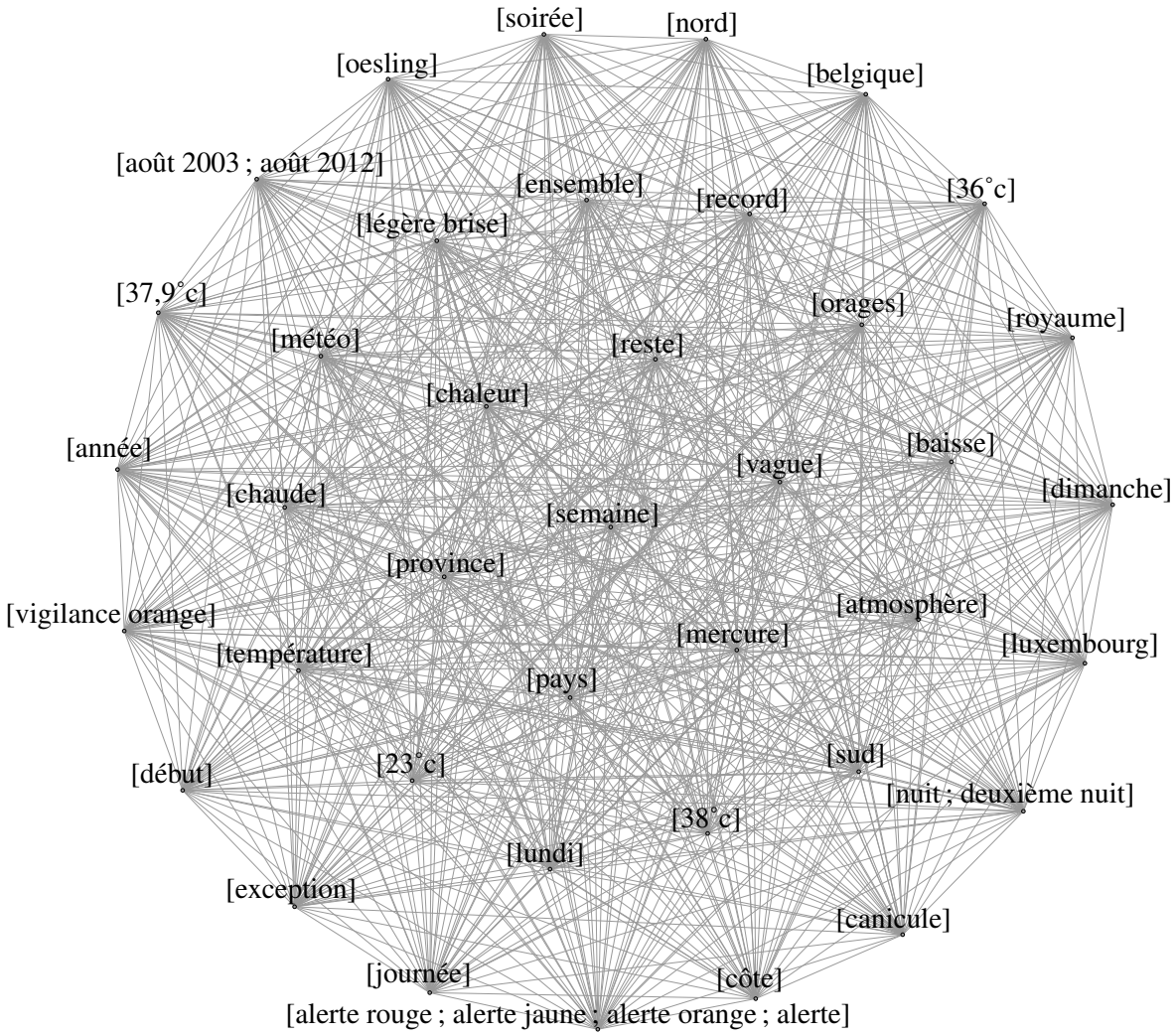
Météo du 19 août 2012 : alerte à la canicule sur la Belgique et le Luxembourg

A l'exception de la province de Luxembourg, en alerte jaune, l'ensemble de la Belgique est en vigilance orange à la canicule. Le Luxembourg n'est pas épargné par la vague du chaleur : le nord du pays est en alerte orange, tandis que le sud a était placé en alerte rouge.

En Belgique, la température n'est pas descendue en dessous des 23˚C cette nuit, ce qui constitue la deuxième nuit la plus chaude jamais enregistrée dans le royaume. Il se pourrait que ce dimanche soit la journée la plus chaude de l'année. Les températures seront comprises entre 33 et 38˚C. Une légère brise de côte pourra faiblement rafraichir l'atmosphère. Des orages de chaleur sont a prévoir dans la soirée et en début de nuit.

Au Luxembourg, le mercure devrait atteindre 32˚C ce dimanche sur l'Oesling et jusqu'à 36˚C sur le sud du pays, et 31 à 32˚C lundi. Une baisse devrait intervenir pour le reste de la semaine. Néanmoins, le record d'août 2003 (37,9˚C) ne devrait pas être atteint.

Termes-clés de référence : luxembourg ; alerte ; météo ; belgique ; août 2012 ; chaleur ; température ; chaude ; canicule ; orange ; la plus chaude.



Sortie de TopicRank : luxembourg ; alerte ; nuit ; belgique ; août 2012 ; chaleur ; température ; chaude ; canicule ; dimanche.

FIGURE 3.2. *Exemple d'extraction de termes-clés avec TopicRank sur un article journalistique de Wikinews. Les termes-clés soulignés sont les termes-clés correctement extraits.*

Méthode	Archéologie			Chimie			Linguistique			Sciences de l'info.		
	P	R	F	P	R	F	P	R	F	P	R	F
TF-IDF	**29,9**	**18,3**	**22**	**14,2**	**12,1**	**12,4**	**11,9**	**13,5**	**12,5**	**12,7**	**13,1**	**12,6**
TextRank	11,9	5,7	7,5	8,9	5,5	6,4	6,9	5,7	6,1	6,6	4,8	5,4
SingleRank	12,3	7,8	9,3	12,4	10,7	10,8	8,7	9,9	9,2	10,8	11,5	10,9
TopicRank	26,6	16,5	19,8	13,5	11,6	11,8	11,1	12,6	11,7	11,7	12,1	11,6

Tableau 3.2. *Résultats de l'extraction de dix mots-clés avec* TF-IDF*, TextRank, SingleRank et TopicRank sur les données de DEFT 2016 en termes de précision (P), rappel (R) et F-mesure (F).*

par une arête pondérée $a \in A$ selon la force du lien sémantique entre les sujets :

$$\text{poids}(n_i, n_j) = \sum_{c_i \in n_i} \sum_{c_j \in n_j} \text{distance}(c_i, c_j) \quad [8]$$

$$\text{distance}(c_i, c_j) = \sum_{p_i \in \text{positions}(c_i)} \sum_{p_j \in \text{positions}(c_j)} \frac{1}{|p_i - p_j|} \quad [9]$$

Plus faible est la distance entre les mots-clés candidats de deux sujets dans le document, plus élevé est le poids de l'arête entre les deux sujets.

4. **Ordonnancement des sujets.** À la manière de TextRank (Mihalcea & Tarau, 2004), TopicRank ordonne les sujets par importance selon le principe de recommandation. Plus un sujet est fortement connecté à un grand nombre de sujets, plus il gagne d'importance, et plus les sujets avec lesquels il est fortement connecté sont importants, plus l'importance qu'il gagne est forte :

$$\text{importance}(n_i) = (1 - \lambda) + \lambda \times \sum_{n_j \in A(n_i)} \frac{\text{poids}(n_i, n_j) \times \text{importance}(n_j)}{\sum_{n_k \in A(n_j)} \text{poids}(n_j, n_k)} \quad [10]$$

Où λ est un facteur de lissage fixé à 0,85 par Brin & Page (1998).

5. **Extraction des mots-clés.** TopicRank extrait un unique mot-clé pour chacun des k plus importants sujets. Bougouin *et al.* (2013) ont choisi de sélectionner dans chaque sujet le mot-clé candidat qui apparaît en premier dans le document.

La figure 3.2 donne un exemple d'extraction de termes-clés avec TopicRank sur un article journalistique de Wikinews. Dans cet exemple, nous observons un groupement correct de toutes les variantes de *alertes*, mais aussi un groupement erroné de *août 2003* avec *août 2012*.

Nous avons appliqué quatre méthodes non-supervisées : TopicRank, TF-IDF, TextRank et SingleRank, sur les données issues de DEFT 2016 pour des mots-clés candidats sélectionnés avec le patron grammatical `/(N|A)+/`. Ce patron est celui utilisé par les expériences de Hasan & Ng (2010) sur des collections de données diverses dont les résultats sont rappelés dans le tableau 2.1. Nous n'avons pas appliqué les méthodes TopicalPageRank et WordTopic-MultiRank présentées en section 2 qui s'appuient sur le modèle d'analyse sémantique latente LDA en fonction d'un nombre de sujets pré-définis observés dans des documents similaires.

Le tableau 3.2 montre les performances de TopicRank comparées à TF-IDF, TextRank et SingleRank. De manière générale, les performances de ces méthodes d'extraction de mots-clés sont basses car, sur ces données, 37 à 76 % de leurs mots-clés n'occurrent pas dans les documents et ne peuvent donc pas être extraits. TopicRank est plus performant que les deux méthodes de référence à base de graphe et confirme donc que le groupement des candidats permet de rassembler des informations pour améliorer la précision de l'ordonnancement. Contrairement à d'autres jeux de données comme celui de DEFT 2012 dédié à l'identification des mots-clés (Bougouin *et al.*, 2013), TopicRank n'est pas plus performant que TF-IDF.

4 TopicCoRank

TopicCoRank (Bougouin *et al.*, 2016) adapte TopicRank et améliore ses performances en tirant partie des éléments du domaine des collections de DEFT 2016. Il réalise donc simultanément deux catégories d'indexation par mots-clés :

— **Extraction de mots-clés** : les mots-clés sont sélectionnés parmi les unités textuelles du document (e.g. TopicRank) ;

— **Assignation de mots-clés** : les mots-clés ne sont pas restreints au contenu du document et doivent faire partie d'un vocabulaire contrôlé construit pour cette tâche.

Pour l'assignation, nous utilisons les mots-clés des notices d'entraînement comme éléments du domaine. Nous ne faisons pas usage des thésaurus fournis par les organisateurs. L'objectif est de proposer une méthode supervisée originale à base de graphe qui ne s'appuie pas sur directement sur un vocabulaire contrôlé, mais exploite les indexations réalisées à l'aide de ce vocabulaire contrôlé.

TopicCoRank modifie les étapes de construction du graphe, d'ordonnancement par importance et de sélection des mots-clés de TopicRank. La construction du graphe étend le graphe de sujet en l'unifiant à un graphe des mots-clés de référence du domaine. L'ordonnancement est désormais conjoint entre les sujets du document et les mots-clés du domaine. Enfin, la sélection des mots-clés ajoute la possibilité de puiser dans le graphe du domaine.

4.1. Construction du graphe

Afin de réaliser simultanément extraction et assignation de mots-clés, TopicCoRank unifie deux graphes : l'un représentant le document (graphe de sujets) et l'autre les mots-clés de référence de son domaine (graphe du domaine). Le premier graphe sert à l'extraction de mots-clés. Le second, construit à partir des mots-clés de référence de documents d'apprentissage, sert à l'assignation. Ce dernier graphe est construit à partir des mots-clés de référence de documents d'entraînement. A l'instar de Chaimongkol & Aizawa (2013) pour l'extraction de termes techniques, nous faisons l'hypothèse que les mots-clés de référence des documents d'apprentissage constituent la terminologie du domaine et nous les utilisons comme substituts au vocabulaire contrôlé usuel en assignation de mots-clés. Contrairement aux mots-clés candidats sélectionnés dans le document, les mots-clés de référence ne sont pas redondants et ne sont donc pas groupés en sujets.

Soit le graphe unifié non orienté $G = (N, A = A_{interne} \cup A_{externe})$. N dénote indifféremment les sujets et les mots-clés du domaine. A regroupe les arêtes $A_{interne}$, qui connectent deux sujets ou deux mots-clés du domaine, et les arêtes $A_{externe}$, qui connectent un sujet à un mot-clé du domaine. La figure 4.3 illustre ce graphe unifié. Deux sujets ou deux mots-clés du domaine sont connectés lorsqu'ils apparaissent dans le même contexte et leur arête est pondérée par le nombre de fois que cela se produit. Lorsqu'il s'agit des sujets, le contexte est une phrase du document ; lorsqu'il s'agit des mots-clés du domaine, le contexte est l'ensemble des mots-clés du domaine d'un document d'apprentissage. Les contextes étant utilisés pour la création du graphe, le graphe de sujets n'est plus complet comme celui de TopicRank.

Le graphe de sujets et le graphe du domaine sont unifiés grâce aux arêtes $A_{externe}$. L'objectif des arêtes $A_{externe}$ est la connexion du document à son domaine par l'intermédiaire des concepts qu'ils partagent. Une arête $A_{externe}$ est donc créée entre un sujet et un mot-clé du domaine si ce dernier appartient au sujet, c'est-à-dire correspond à l'un de ses mots-clés candidats.

4.2. Ordonnancement conjoint des sujets et des mots-clés du domaine

L'ordonnancement conjoint des sujets et des mots-clés du domaine établit leur ordre d'importance vis-à-vis du contenu du document et du domaine. Pour cela, un score d'importance est attribué simultanément aux sujets et aux mots-clés du domaine. Nous reprenons le principe de la recommandation de TopicRank et l'adaptons au problème d'ordonnancement conjoint. Les premières hypothèses de recommandation sont donc les mêmes que celle de TopicRank :

— un sujet est d'autant plus important s'il est fortement connecté à un grand nombre de sujets et si les sujets avec lesquels il est fortement connecté sont importants ;

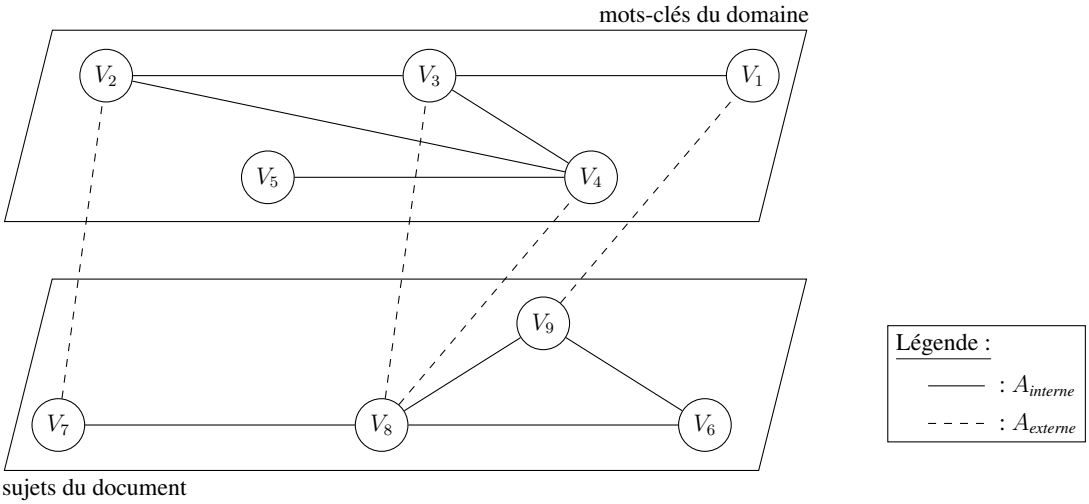


FIGURE 4.3. *Illustration du graphe unifié utilisé par TopicCoRank.*

λ_s	λ_m								
	0,10	0,20	0,30	0,40	0,50	0,60	0,70	0,80	0,90
0,10	0,265929	0,223061	0,220135	0,210187	0,205136	0,200050	0,182961	0,163289	0,152953
0,20	0,308469	0,268461	0,235401	0,220135	0,208541	0,200926	0,180708	0,163300	0,145991
0,30	0,328994	0,299389	0,268872	0,239055	0,216093	0,198867	0,175048	0,159956	0,140896
0,40	0,356845	0,329772	0,300162	0,272695	0,229187	0,199256	0,170756	0,149463	0,128656
0,50	**0,369872**	0,357373	0,332364	0,302425	0,252934	0,201961	0,168769	0,138578	0,114714
0,60	0,329932	0,316328	0,304142	0,287540	0,273105	0,232559	0,172237	0,130202	0,099038
0,70	0,190299	0,190299	0,194126	0,197638	0,205074	0,209377	0,199643	0,117608	0,089627
0,80	0,175659	0,176659	0,176940	0,178260	0,178260	0,178260	0,183054	0,184146	0,086289
0,90	0,170923	0,170923	0,170871	0,171871	0,171994	0,171994	0,171994	0,169994	0,180775

Tableau 4.3. *F-mesure de TopicCoRank pour la linguistique selon les valeurs λ_s et λ_m*

— un mot-clé du domaine est d'autant plus important s'il est fortement connecté à un grand nombre de mots-clés du domaine et si les mots-clés du domaine avec lesquels il est connecté sont importants.

Ces hypothèses de recommandation proposent une recommandation interne : elles établissent l'importance des sujets les uns par rapport aux autres et l'importance des mots-clés du domaine les uns par rapport aux autres. Cependant, elles ne permettent pas de tirer profit des relations entre sujets et mots-clés du domaine. Par ailleurs, l'importance des mots-clés du domaine ne dépend pas du document. Nous ajoutons donc deux nouvelles hypothèses de recommandation, que nous qualifions d'externes :

— un sujet est d'autant plus important s'il est représenté par (connecté à) des mots-clés du domaine importants ;
— un mot-clé du domaine est d'autant plus important vis-à-vis du contenu du document s'il véhicule (est connecté à) l'un de ses sujets importants.

Sujets et mots-clés du domaine sont ainsi évalués d'après leur usage dans le document et leur importance dans le domaine. L'ordonnancement des uns joue un rôle sur celui des autres et permet ainsi d'effectuer conjointement extraction et assignation.

À partir du graphe unifié, nous ordonnons simultanément sujets $s \in N$ du document et mots-clés $m \in N$ du domaine par importance. Pour cela, nous reprenons le même principe que TopicRank et l'adaptons de sorte que sujets et mots-clés du domaine se transfèrent de l'importance. Nous proposons deux formes de recommandation : une recommandation interne $R_{interne}$ qui intervient entre deux nœuds du même type (sujets ou mots-clés du

λ_k	λ_t								
	0,10	0,20	0,30	0,40	0,50	0,60	0,70	0,80	0,90
0,10	0,287552	0,263499	0,258782	0,256730	0,251521	0,241996	0,233082	0,228069	0,207586
0,20	0,315930	0,288671	0,268896	0,261621	0,260268	0,252238	0,239033	0,225557	0,199013
0,30	0,332686	0,308754	0,288124	0,266357	0,261329	0,252295	0,239102	0,208736	0,193353
0,40	0,346112	0,328580	0,304048	0,283733	0,260291	0,252718	0,232835	0,198415	0,185377
0,50	**0,370700**	0,353686	0,330340	0,298682	0,267274	0,246244	0,225128	0,192599	0,171164
0,60	0,333296	0,324421	0,322860	0,310443	0,282135	0,246035	0,217478	0,176974	0,158986
0,70	0,193747	0,178799	0,174778	0,169829	0,175432	0,202766	0,216087	0,163864	0,140023
0,80	0,097146	0,097183	0,095015	0,091848	0,090515	0,096530	0,113824	0,162864	0,131422
0,90	0,073907	0,071554	0,071554	0,070645	0,070645	0,070645	0,070645	0,070645	0,131578

Tableau 4.4. *F-mesures de TopicCoRank pour les sciences de l'information selon les valeurs de λ_s and λ_m*

λ_k	λ_t								
	0,10	0,20	0,30	0,40	0,50	0,60	0,70	0,80	0,90
0,10	0,413157	0,393052	0,373521	0,364198	0,352233	0,333856	0,317894	0,283662	0,256224
0,20	0,440457	0,410872	0,393465	0,384026	0,363234	0,338660	0,312442	0,278358	0,237849
0,30	0,454136	0,420689	0,404031	0,389898	0,366015	0,338598	0,304129	0,260972	0,227070
0,40	0,454808	0,439736	0,414329	0,395066	0,365365	0,329143	0,291480	0,241169	0,219739
0,50	**0,468462**	0,465186	0,444262	0,410748	0,369891	0,320035	0,275065	0,226923	0,199869
0,60	0,408617	0,418801	0,405849	0,394795	0,375971	0,340019	0,270603	0,213206	0,177203
0,70	0,253766	0,253212	0,253855	0,256144	0,270933	0,299230	0,290054	0,207877	0,161434
0,80	0,209198	0,209768	0,208443	0,207140	0,208652	0,215149	0,218337	0,254218	0,155159
0,90	0,197547	0,197547	0,199225	0,200134	0,200848	0,200848	0,201589	0,202472	0,233327

Tableau 4.5. *F-mesures de TopicCoRank pour l'archéologie selon les valeurs de λ_s and λ_m*

domaine) et une recommandation externe $R_{externe}$ qui intervient entre un sujet et un mot-clé du domaine.

$$\text{importance}(s_i) = (1 - \lambda_s)\ R_{externe}(s_i) + \lambda_s\ R_{interne}(s_i) \quad [11]$$

$$\text{importance}(m_i) = (1 - \lambda_m)\ R_{externe}(m_i) + \lambda_m\ R_{interne}(m_i) \quad [12]$$

$$R_{interne}(n_i) = \sum_{n_j \in A_{interne}(n_i)} \frac{\text{poids}(n_i, n_j)(n_j)}{\sum_{n_k \in A(n_j)} \text{poids}(n_j, n_k)} \quad [13]$$

$$R_{externe}(n_i) = \sum_{n_j \in A_{externe}(n_i)} \frac{\text{importance}(n_j)}{|A_{out}(n_j)|} \quad [14]$$

λ_s et λ_m sont des paramètres permettant de contrôler l'influence de la recommandation interne sur la recommandation externe respectivement pour les sujets et pour les mots-clés du domaine avec $0 \leq \lambda_s \leq 1$ et $0 \leq \lambda_m \leq 1$. Une valeur importante de λ_s ou de λ_m donne plus d'influence à la recommandation interne. Les valeurs de λ_s et λ_m ont été estimées sur les corpus d'apprentissage. Les tableaux 4.3, 4.4 et 4.5 précisent pour chacun des domaines les scores de F-mesure obtenus selon les valeurs de λ_s et de λ_m. Les meilleures valeurs de F-mesure étant obtenues pour λ_m à 0,1 et pour λ_s à 0,5 quel que soit le domaine, nous retenons ces valeurs pour calculer l'importance respective des sujets et des mots-clés du domaine.

4.3. Sélection des mots-clés

TopicCoRank utilise l'ordre d'importance établit avec le score S des sujets et mots-clés du domaine pour déterminer les mots-clés du document. Les k nœuds du graphe unifié ayant obtenu les meilleurs scores sont retenus, qu'ils soient des sujets ou des mots-clés du domaine.

Méthode	Archéologie			Chimie			Linguistique			Sciences de l'information		
	P		F	P	R	F	P	R	F	P	R	F
TopicCoRank	49,86	31,16	37,28	20,87	17,45	18,11	22,23	24,87	23,24	20,61	20,65	20,21
TopicCoRank$_{extr.}$	43,63	26,63	32,17	15,77	13,10	13,60	13,77	15,56	14,47	15,67	15,87	15,39
TopicCoRank$_{assign.}$	**53,77**	**33,46**	**40,11**	**21,15**	**17,54**	**18,28**	**23,16**	**25,85**	**24,19**	**21,93**	**21,83**	**21,45**

Tableau 6.6. *Résultats des trois exécutions de TopicCoRank soumises à DEFT 2016 pour les collections d'archéologie, de chimie, de linguistique et de sciences de l'information en termes de précision (P), rappel (R) et F-mesure (F).*

Lorsqu'un mot-clé du domaine doit être assigné, une étape de vérification supplémentaire est effectuée pour s'assurer de son importance à la fois pour le domaine et pour le document. En effet, il est possible que le graphe du domaine soit constitué de composantes connexes où les nœuds représentant des mots-clés du domaine ne sont connectés qu'entre eux sans l'être à un sujet du document. Si son importance est déterminée uniquement à partir du domaine, il n'est pas souhaitable d'assigner ce mot-clé au document.

Lorsqu'un nœud retenu représente un sujet, c'est la même stratégie que celle de TopicRank qui est appliquée. Pour un sujet donné, le mot-clé extrait est son mot-clé candidat qui apparaît en premier dans le document.

La figure 4.4 donne un exemple d'extraction et d'assignation des mots-clés avec TopicCoRank à partir d'une notice d'archéologie. Dans cet exemple, nous observons une meilleure indexation par mots-clés qu'avec TopicRank. Tout d'abord, nous voyons que le graphe du domaine permet l'assignation du mot-clé générique « France ». Ensuite, nous voyons que les relations de « diffusion », « analyse » et « répartition » dans le graphe du domaine permettent de mieux les ordonner.
TopicCorank applique un algorithme de regroupement hiérachique agglomératif de complexité $n^2log(n)$, puis un parcours de graphes effectué en temps linéaire.

5 Méthodes évaluées

Nous avons soumis trois méthodes réalisant un ordonnancement conjoint : TopicCoRank et deux variantes, l'une effectuant uniquement une extraction, TopicCoRank$_{extr.}$, et l'autre uniquement une assignation, TopicCoRank$_{assign.}$. TopicCoRank$_{extr.}$, la variante d'extraction de TopicCoRank, propose des mots-clés appartenant aux sujets. Le graphe unifié est utilisé ainsi que les mots-clés du domaine pour l'ordonnancement.
TopicCoRank$_{assign.}$, la variante d'assignation de TopicCoRank, propose des mots-clés du domaine. Le graphe unifié est utilisé ainsi que les sujets du document pour l'ordonnancement.

Pour ces trois méthodes, nous proposons dix mots-clés parmi les sujets et/ou mots-clés du domaine les plus importants quelle que soit la notice et quel que soit le domaine.

6 Résultats

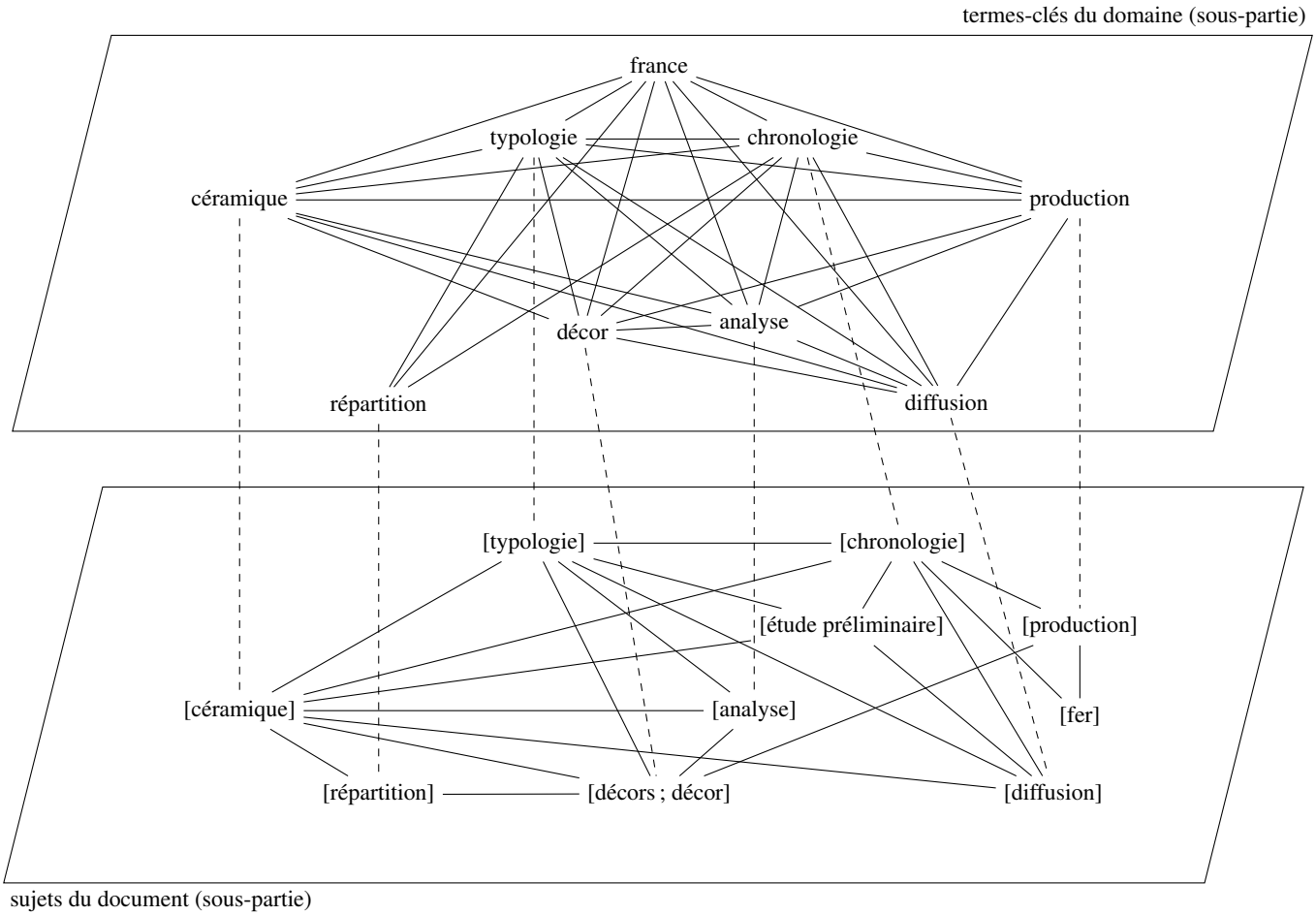
Nous présentons dans le tableau 6.6 les résultats officiels de la campagne DEFT 2016 pour TopicCoRank et les deux variantes d'extraction TopicCoRank$_{extr.}$ et d'assignation TopicCoRank$_{assign.}$ pour les collections d'archéologie, de chimie, de linguistique et de sciences de l'information.

Les résultats obtenus sur le corpus de test par TopicCoRank comparés avec ceux de TopicRank (cf. Tableau 3.2) montrent une augmentation de 6 points de F-mesure pour la chimie, de 9 points pour les sciences de l'information, de 11 points pour la linguistique, et de 17 points pour l'archéologie. Ils démontrent l'efficacité de la méthode d'apprentissage modélisée dans TopicCoRank tirant partie des notices déjà indexées.
Parmi les trois variantes, c'est la variante d'assignation de TopicCoRank$_{assign.}$ qui obtient les meilleurs résultats. Les résultats sont améliorés par rapport à TopicCoRank d'un point de F-mesure pour la linguistique et les sciences de l'information et de 3 points pour la chimie. Les mots-clés du domaine gagnent donc à être ordonnés d'après le contenu du document (ses sujets). Compte tenu de la nature des collections de données, cette observation semble normale. En effet, les notices sont indexées par des indexeurs professionnels utilisant principalement un

Étude préliminaire de la céramique non tournée micacée du bas Languedoc occidental : typologie, chronologie et aire de diffusion

L'étude présente une variété de céramique non tournée dont la typologie et l'analyse des décors permettent de l'identifier facilement. La nature de l'argile enrichie de mica donne un aspect pailleté à la pâte sur laquelle le décor effectué selon la méthode du brunissoir apparaît en traits brillant sur fond mat. Cette première approche se fonde sur deux séries issues de fouilles anciennes menées sur les oppidums du Cayla à Mailhac (Aude) et de Mourrel-Ferrat à Olonzac (Hérault). La carte de répartition fait état d'échanges ou de commerce à l'échelon macrorégional rarement mis en évidence pour de la céramique non tournée. S'il est difficile de statuer sur l'origine des décors, il semble que la production s'insère dans une ambiance celtisante. La chronologie de cette production se situe dans le deuxième âge du Fer. La fourchette proposée entre la fin du IV^{e} et la fin du II^{e} s. av. J.-C. reste encore à préciser.

Termes-clés de référence : distribution ; mourrel-ferrat ; olonzac ; le cayla ; mailhac ; micassé ; céramique non-tournée ; celtes ; production ; echange ; commerce ; cartographie ; habitat ; oppidum ; site fortifié ; fouille ancienne ; identification ; décor ; analyse ; répartition ; diffusion ; chronologie ; typologie ; céramique ; etude du matériel ; hérault ; aude ; france ; europe ; la tène ; age du fer.



Sortie de TopicCoRank : céramique ; décors ; typologie ; chronologie ; production ; étude préliminaire ; diffusion ; analyse ; france ; répartition.

Sortie de TopicRank : décors ; céramique ; chronologie ; typologie ; production ; fin ; étude préliminaire ; fer ; deuxième âge ; aire.

FIGURE 4.4. *Exemple d'extraction de termes-clés avec TopicCoRank sur le résumé d'une notice d'archéologie. Les mots-clés soulignés sont les mots-clés correctement extraits.*

	Extraction (%)	Assignation (%)
Archéologie *(fr)*	69,1	30,9
Chimie *(fr)*	68,4	31,6
Linguistique *(fr)*	61,7	38,3
Sciences de l'info. *(fr)*	66,4	33,6

Tableau 6.7. *Taux moyens d'extraction et d'assignation obtenus avec TopicCoRank sur la totalité des données DEFT 2016 (apprentissage et test).*

vocabulaire contrôlé. Néanmoins, TopicCoRank ne fait pas suffisamment émerger les mots-clés du domaine. Afin d'observer la place que prend l'assignation dans TopicCoRank, et pour comprendre pourquoi sa variante $\text{TopicCoRank}_{\text{assign.}}$ est plus performante, nous nous intéressons maintenant aux taux de mots-clés extraits et assignés par TopicCoRank, présentés dans le tableau 6.7. Nous observons que l'extraction est légèrement prédominante face à l'assignation. Les deux catégories d'indexation par mots-clés sont effectivement réalisées conjointement, mais l'ordonnancement donne plus d'importance aux sujets du document qu'aux mots-clés de référence du domaine. En domaines de spécialité où l'assignation est préférée, cela peut être résolu en travaillant sur un affinage des schémas de connexion des nœuds de chaque graphe et d'unification de ceux-ci.

	Archéologie		Chimie		Linguistique		Sciences de l'information	
Rang	Équipe	F-mesure	Équipe	F-mesure	Équipe	F-mesure	Équipe	F-mesure
1	EXENSA	45,59	EXENSA	21,46	EBSIUM	31,75	EBSIUM	28,98,
2	LIMSI	43,26	EBSIUM	21,07	EXENSA	26,30	EXENSA	23,86
3	**LINA**	**40,11**	**LINA**	**18,28**	**LINA**	**24,19**	**LINA**	**21,45**
4	EBSIUM	34,96	LIPN	15,31	LIPN	19,07	LIPN	15,34
5	LIPN	30,75	LIMSI	15,29	LIMSI	15,63	LIMSI	12,49

Tableau 6.8. *Classement de DEFT 2016 sur la base des meilleures soumissions pour les collections d'archéologie, de chimie, de linguistique et de sciences de l'information. Notre classement est indiqué en gras.*

Le tableau 6.8 présente, pour les collections d'archéologie, de chimie, de linguistique et de sciences de l'information, le classement des différentes équipes sur la base de la meilleure soumission. Notre soumission est classée au rang 3 sur 5.

7 Analyse des mots-clés proposés par TopicCoRank

Dans cette section, nous analysons quelques mots-clés corrects (vrais positifs) et incorrects (faux positifs) issus du graphe du domaine des collections DEFT 2016.

7.1. Analyse des vrais positifs

Parmi les mots-clés assignés, ceux qui sont corrects sont en grande partie présents dans le contenu du document. Ils sont directement connectés aux sujets du document et leur importance respective évolue de manière similaire. Il est fréquent qu'un mot-clé candidat d'un sujet soit extrait en même temps qu'un mot-clé du domaine connecté à ce sujet. Dans cette situation, un seul mot-clé est conservé si le mot-clé extrait et celui assigné sont identiques, sinon les deux mots-clés sont conservés.

Les mots-clés du domaine corrects, absents du document, qui sont connectés indirectement aux sujets du document sont rarement retenus. Seuls le terme *analyse du discours* en linguistique, les noms de composés comme *composé aliphatique* ou *composé benzénique* en chimie, ou des mots-clés fréquemment employés comme *français* en linguistique, ou encore *Europe* en archéologie ont été sélectionnés (*français* apparaît dans 48,9 % des documents de linguistique et *Europe* apparaît dans 52,5 % des documents d'archéologie).

7.2. Analyse des faux positifs

Les mots-clés génériques évoqués dans l'analyse des vrais positifs sont aussi sources d'erreurs. En effet, comme ils sont associés à un nombre conséquent de documents d'entraînement, ils sont connectés à beaucoup d'autres mots-clés du domaine et gagnent donc de l'importance quelque soit le document. Pour l'exemple du mot-clé *français* en linguistique, nous observons des documents qui traitent de l'arabe mais, parce que les mots techniques employés sont les mêmes (*syntaxe*, *sémantique*, etc.), *français* est assigné.

Enfin, nous observons quelques problèmes liés à la présence de mots-clés de référence redondants, c'est-à-dire des synonymes. C'est le cas, par exemple, du mot-clé *pratique funéraire* parfois utilisé au lieu du mot-clé *rite funéraire* appartenant au vocabulaire contrôlé d'archéologie. L'évaluation considère comme une erreur un synonyme d'un mot-clé de référence.

8 Conclusion

Nous avons décrit la participation du LINA à DEFT 2016. Nous avons proposé une méthode à base de graphe, TopicCoRank, qui est une extension de la méthode d'extraction de mots-clés TopicRank et qui utilise les mots-clés de référence des documents d'entraînement comme vocabulaire contrôlé. TopicCoRank combine deux graphes représentant le document à analyser et son domaine de spécialité, lui permettant d'extraire des mots-clés du document et d'en assigner à partir de son domaine. Ces derniers n'apparaissent pas nécessairement dans le document. Notre système s'est classé à la troisième place sur un total de cinq systèmes.
Parmi les trois variantes de TopicCorank que nous avons proposées, celle qui ne réalise que l'assignation de mots-clés est la meilleure. Cela signifie que notre modèle hybride pour l'extraction et l'assignation simultanés de mots-clés nécessite d'être améliorer. Nous envisageons d'étendre ce travail en affinant le schéma de connection des deux graphes afin de faciliter l'émergence des mots-clés à assigner.

9 Remerciements

Ce travail a bénéficié d'une aide de l'Agence Nationale de la Recherche portant la référence (ANR-12-CORD-0029).

Références

ADAM J., BARON D., EDWARDS J., ELLIS D., GELBART D., MORGAN N., PESKIN B., PFAU T., SHRIBERG E., STOLCKE A. & WOOTERS C. (2003). The ICSI Meeting Corpus. In *Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03). 2003 IEEE International Conference on*, volume 1, p. I–364–I–367 vol.1.

BLEI D. M., NG A. Y. & JORDAN M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3, 993–1022.

BOUGOUIN A. & BOUDIN F. (2014). TopicRank : ordonnancement de sujets pour l'extraction automatique de termes-clés. *TAL*, 55(1), 45–69.

BOUGOUIN A., BOUDIN F. & DAILLE B. (2013). Topicrank : Graph-Based Topic Ranking for Keyphrase Extraction. In *Proceedings of the 6th International Joint Conference on Natural Language Processing (IJCNLP)*, p. 543–551, Nagoya, Japan : Asian Federation of Natural Language Processing.

BOUGOUIN A., BOUDIN F. & DAILLE B. (2016). Keyphrase annotation with graph co-ranking. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics : Technical Papers*, p. 2945–2955, Osaka, Japan : The COLING 2016 Organizing Committee.

BRIN S. & PAGE L. (1998). The Anatomy of a Large-Scale Hypertextual Web Search Engine. *Computer Networks and ISDN Systems*, 30(1), 107–117.

CHAIMONGKOL P. & AIZAWA A. (2013). Utilizing LDA Clustering for Technical Term Extraction. In *Proceedings of the 19th Annual Meeting of the Association for Natural Language Processing (ANLP)*, p. 686–689, Nagoya, Japan : Association for Natural Language Processing.

HASAN K. S. & NG V. (2010). Conundrums in Unsupervised Keyphrase Extraction : Making Sense of the State-of-the-Art. In *Proceedings of the 23rd International Conference on Computational Linguistics : Posters (COLING)*, p. 365–373, Stroudsburg, PA, USA : Association for Computational Linguistics.

HULTH A. (2003). Improved Automatic Keyword Extraction Given More Linguistic Knowledge. In *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, p. 216–223, Stroudsburg, PA, USA : Association for Computational Linguistics.

Z. KOZAREVA, I. MATVEEVA, G. MELLI & V. NASTASE, Eds. (2013). *Proceedings of TextGraphs-8 Graph-Based Methods for Natural Language Processing*. Seattle, Washington, USA : Association for Computational Linguistics.

LIU K., XU L. & ZHAO J. (2014). Extracting Opinion Targets and Opinion Words from Online Reviews with Graph Co-ranking. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 314–324, Baltimore, Maryland : Association for Computational Linguistics.

LIU Z., HUANG W., ZHENG Y. & SUN M. (2010). Automatic Keyphrase Extraction Via Topic Decomposition. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing(EMNLP)*, p. 366–376, Stroudsburg, PA, USA : Association for Computational Linguistics.

MIHALCEA R. & TARAU P. (2004). TextRank : Bringing Order Into Texts. In DEKANG LIN & DEKAI WU, Eds., *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, p. 404–411, Barcelona, Spain : Association for Computational Linguistics.

NGUYEN T. D. & KAN M.-Y. (2007). Keyphrase Extraction in Scientific Publications. In *Proceedings of the 10th International Conference on Asian Digital Libraries : Looking Back 10 Years and Forging New Frontiers*, p. 317–326, Berlin, Heidelberg : Springer-Verlag.

PORTER M. F. (1980). An Algorithm for Suffix Stripping. *Program : Electronic Library and Information Systems*, 14(3), 130–137.

SALTON G., WONG A. & YANG C. (1975). A Vector Space Model for Automatic Indexing. *Communication ACM*, 18(11), 613–620.

WAN X. (2011). Using Bilingual Information for Cross-Language Document Summarization. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics : Human Language Technologies-Volume 1*, p. 1546–1555 : Association for Computational Linguistics.

WAN X. & XIAO J. (2008). Single Document Keyphrase Extraction Using Neighborhood Knowledge. In *Proceedings of the 23rd National Conference on Artificial Intelligence - Volume 2*, p. 855–860 : AAAI Press.

YAN R., LAPATA M. & LI X. (2012). Tweet Recommendation with Graph Co-Ranking. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 516–525, Jeju Island, Korea : Association for Computational Linguistics.

ZHANG F., HUANG L. & PENG B. (2013). WordTopic-MultiRank : A New Method for Automatic Keyphrase Extraction. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, p. 10–18, Nagoya, Japan : Asian Federation of Natural Language Processing.

