TopicRank : Ordonnancement de sujets pour l'extraction automatique de termes-clés

Adrien Bougouin* — **Florian Boudin***

* *LINA - UMR CNRS 6241, Université de Nantes*
UFR de Sciences et Techniques, 2 rue de la Houssinière, 44322 Nantes, France
prenom.nom@univ-nantes.fr

RÉSUMÉ. *Les termes-clés sont les mots ou expressions polylexicales qui représentent les informations essentielles d'un document. Du fait de leur utilité pour diverses applications, telles que l'indexation automatique ou le résumé automatique, mais aussi parce que la quantité de données aujourd'hui accessibles est trop importante pour être traitée manuellement, la tâche d'extraction automatique de termes-clés suscite de plus en plus l'intéret des chercheurs. Dans cet article nous présentons TopicRank, une méthode non-supervisée à base de graphe pour l'extraction de termes-clés. Cette méthode regroupe les termes-clés candidats en sujets, ordonne les sujets par importance, et extrait le candidat le plus représentatif de chacun des meilleurs sujets. Les expériences réalisées montrent une amélioration significative vis-à-vis de l'état de l'art des méthodes à base de graphe pour l'extraction de termes-clés.*

ABSTRACT. *Keyphrases are single or multi-word expressions that represent the main content of a document. As keyphrases are useful in many applications such as document indexing or text summarization, and also because the vast amount of data available nowadays can not be manually annotated, the task of automatically extracting keyphrases has attracted considerable attention. In this article we present TopicRank, an unsupervised graph-based method for keyphrase extraction. This method clusters the keyphrase candidates into topics, ranks these topics and extracts the most representative candidate for each of the best topics. Our experiments show a significant improvement over the state-of-the-art graph based methods for keyphrase extraction.*

MOTS-CLÉS : *extraction de termes-clés, groupement en sujets, ordonnancement de sujets, méthode non-supervisée, méthode à base de graphe*

KEYWORDS: *keyphrase extraction, topic clustering, topic ranking, unsupervised method, graph-based method*

1re soumission à *TAL 55-1*, le 15/11/2014

1. Introduction

Les termes-clés sont des mots ou des expressions polylexicales qui représentent les sujets principaux du document auquel ils se réfèrent. Du fait de leur propriété synthétique, les termes-clés sont utilisés dans de nombreuses applications telles que l'indexation de documents (Medelyan et Witten, 2008), le résumé automatique (Litvak et Last, 2008) ou encore la classification de documents (Han *et al.*, 2007). Avec la croissance exponentielle du nombre de documents disponibles, en particulier sur le web, les termes-clés constituent un moyen efficace pour accéder rapidement aux informations pertinentes. Cependant, la plupart des documents ne sont pas pourvus de termes-clés et l'annotation manuelle de ces derniers est une tâche beaucoup trop coûteuse pour être envisagée. C'est la raison pour laquelle la problématique de l'extraction automatique de termes-clés suscite de plus en plus l'intéret des chercheurs.

L'extraction automatique de termes-clés peut être perçue soit comme une tâche de classification binaire de termes-clés candidats (Witten *et al.*, 1999), soit comme une tâche d'ordonnancement de termes-clés candidats (Mihalcea et Tarau, 2004). Dans le premier cas, l'extraction est le plus souvent supervisée, elle nécessite une collection de documents annotés en termes-clés pour une phase d'apprentissage préliminaire. Dans le second cas, l'extraction est le plus souvent non-supervisée, elle ne nécessite pas de documents préalablement annotés. Les méthodes non-supervisées ont des performances plus faibles que les méthodes supervisées actuelles. Ceci s'explique par le fait que les méthodes supervisées apprennent les critères discriminants pour l'extraction de termes-clés à partir de documents annotés. Cependant, la dépendance de ces méthodes au domaine des documents utilisés lors de l'apprentissage pousse de plus en plus de chercheurs à s'intéresser à l'extraction non-supervisée de termes-clés.

Dans cet article, nous présentons TopicRank, une méthode d'extraction non-supervisée de termes-clés qui se fonde sur les travaux de Mihalcea et Tarau (2004, TextRank) pour l'ordonnancement à base de graphe des unités textuelles d'un document. TopicRank groupe les termes-clés candidats selon leur appartenance à un sujet, ordonne les sujets par importance dans le document, puis sélectionne pour chacun des meilleurs sujets le candidat qui le représente le mieux (son terme-clé associé). Contrairement à l'ordonnancement des mots tel qu'il est fait avec TextRank, le groupement des candidats en sujets utilisés pour l'ordonnancement permet de tirer partie d'informations complémentaires extraites de différents candidats d'un même sujet. De plus, le fait de ne sélectionner qu'un seul candidat par sujet permet d'éviter l'extraction de termes-clés redondants.

Pour évaluer TopicRank, nous utilisons quatre collections de données dont les propriétés diffèrent (nature, langue, taille des documents, etc.) afin de mieux observer ses avantages et ses faiblesses (Hasan et Ng, 2010). En addition, nous comparons TopicRank avec trois méthodes non-supervisées, l'une extrayant des statistiques à partir de documents supplémentaires et les deux autres appartenant à la catégorie des méthodes à base de graphe. Pour trois des collections utilisées, TopicRank donne de meilleurs

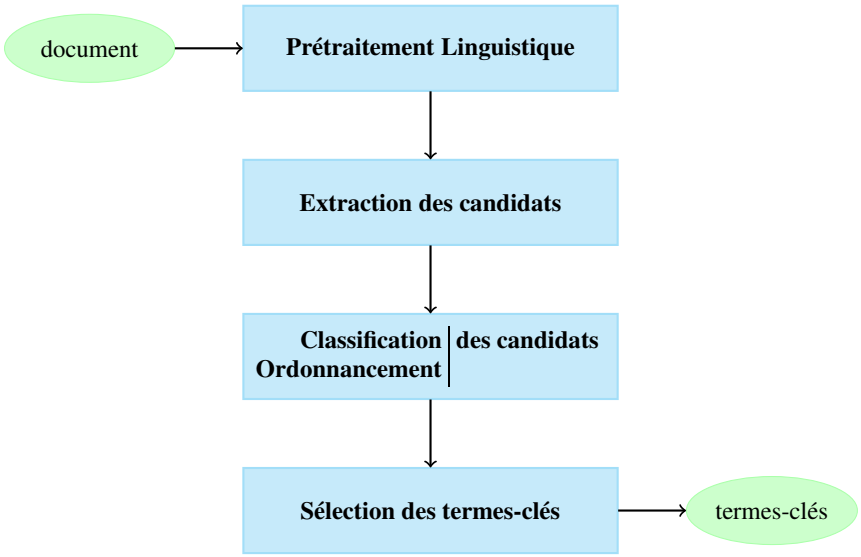


Figure 1. *Les quatre principales étapes de l'extraction automatique de termes-clés.*

résultats que les autres méthodes. De plus, cette amélioration est significative vis-à-vis des deux méthodes à base de graphe.

L'article est structuré comme suit. Après un état de l'art des méthodes d'extraction automatique de termes-clés en section 2, nous décrivons le fonctionnement de TopicRank en section 3 et présentons son évaluation approfondie en section 4. Enfin, nous concluons et discutons des futurs travaux dans la section 5.

2. État de l'art

L'extraction automatique de termes-clés est une tâche répartie en quatre étapes. Les documents sont traités un par un. Ils sont tout d'abord enrichis linguistiquement (segmentés en phrases, segmentés en mots, étiquetés en parties du discours, etc.), puis des termes-clés candidats en sont extraits et classifiés, ou ordonnés, afin de pouvoir sélectionner leurs termes-clés (cf. figure 1). L'extraction des termes-clés candidats et leur classification/ordonnancement sont les deux étapes auxquels nous nous intéressons dans cet article. En effet, la classification/l'ordonnancement des termes-clés candidats est le cœur de la tâche d'extraction de termes-clés et ses performances dépendent de la qualité des candidats préalablement extraits.

2.1. *Extraction de termes-clés candidats*

L'objectif de l'extraction de termes-clés candidats est de réduire l'espace des solutions possibles aux seules unités textuelles ayant des particularités semblables à celles des termes-clés tels qu'ils peuvent être donnés par des humains. Deux avantages directs à cela sont la réduction du temps de calcul nécessaire et la suppression d'unités textuelles non pertinentes pouvant engendrer du bruit affectant les performances de l'extraction de termes-clés. Pour distinguer les différents candidats extraits, nous définissons deux catégories : les candidats positifs, qui sont de réels termes-clés, et les candidats non positifs, qui ne sont pas de réels termes-clés. Parmi les candidats non positifs, nous distinguons aussi deux catégories : les candidats porteurs d'informations utiles à la promotions de candidats positifs et les candidats non pertinents, qui sont considérés comme des erreurs d'extraction.

Dans les travaux précédents concernant l'extraction automatique de termes-clés, trois méthodes d'extraction de candidats sont classiquement utilisées : l'extraction de n-grammes filtrés avec une liste de mots outils, l'extraction d'unités minimales de sens ayant pour tête un nom (chunks nominaux) ou l'extraction des unités textuelles respectant certains patrons syntaxiques.

L'extraction de n-grammes consiste en l'extraction de séquences ordonnées de n mots. Cette extraction est très exhaustive, elle fournit une grande quantité de termes-clés candidats, maximisant ainsi la quantité de candidats positifs ou porteurs d'informations, mais aussi la quantité de candidats non pertinents. Pour supprimer un grand nombre de candidats non pertinents, il est courant d'utiliser une liste de mots outils (conjonctions, prépositions, mots usuels, etc.). Une unité textuelle contenant un mot outils au début ou à la fin ne doit pas être considérée comme un terme-clé candidat. Bien que l'extraction de n-grammes filtrés fournisse un ensemble bruité de candidats, elle est encore largement utilisée parmi les méthodes supervisées d'extraction automatique de termes-clés (Witten *et al.*, 1999 ; Turney, 2000 ; Hulth, 2003). Ceci est dû au fait que leur phase d'apprentissage les rend plus robustes et donc moins sensibles aux bruits que les autres méthodes.

L'extraction de chunks nominaux consiste en l'extraction d'unités minimales de sens ayant pour tête un nom. Contrairement aux n-grammes, les chunks nominaux sont toujours des unités textuelles grammaticalement correctes. D'un point de vue linguistique, l'extraction de chunks nominaux est donc plus justifiée que l'extraction de n-grammes filtrés. Cependant son caractère plus restrictif ne permet pas d'extraire autant de candidats positifs qu'avec les n-grammes. Il est donc important de s'assurer que les propriétés des chunks nominaux sont en accord avec les propriétés des termes-clés tels qu'ils peuvent être donnés par des humains. Les expériences menées par Hulth (2003) et Eichler et Neumann (2010) avec les chunks nominaux montrent une amélioration des performances vis-à-vis de l'usage des n-grammes. Cependant, Hulth (2003) montre aussi qu'en tirant partie de l'étiquetage en parties du discours des termes-clés candidats, l'extraction supervisée de termes-clés à partir de n-grammes donne des performances au-dessus de celles obtenues avec les chunks nominaux.

L'extraction d'unités textuelles respectant certains patrons syntaxiques permet l'extraction de candidats qui sont syntaxiquement contrôlés. Du fait de la syntaxe contrôlée, cette extraction est, tout comme l'extraction de chunks nominaux, plus fondée linguistiquement que la simple extraction de n-grammes filtrés. Alors que Hulth (2003) extraient des candidats avec les patrons de termes-clés les plus fréquents (plus de 10 occurrences) dans une collection de documents annotés, d'autres chercheurs tels que Wan et Xiao (2008) se concentrent uniquement sur les plus longues séquences de noms (noms propres inclus) et d'adjectifs. Pour des méthodes non-supervisées telles que la notre, l'extraction des séquences de noms et d'adjectifs est intéressante, car elle nécessite ni des données supplémentaires, ni une adaptation particulière pour une langue donnée, tel que c'est le cas pour l'extraction des chunks nominaux.

Dans le but d'améliorer la qualité des candidats extraits à partir d'articles scientifique, Kim *et al.* (2009) proposent un filtrage des candidats en fonction de leur spécificité vis-à-vis du document analysé. Cette spécificité est déterminée en fonction du rapport entre la fréquence d'un candidat dans le document et le nombre de documents, d'une collection, dans lesquels il est présent (Spärck Jones, 1972, TF-IDF). Intuitivement, un candidat très fréquent dans le document analysé est d'autant plus spécifique à celui-ci s'il est présent dans très peu d'autres documents. Ce type d'approche est intéressant, mais cela suppose que les documents traitées aient une forte cohérence de domaine. Les documents de nos collections de données ne répondent pas tous à ce critère, nous ne pouvons donc pas appliquer cette méthode, qui en plus requière des documents supplémentaires et la définition d'un seuil pour le filtrage.

2.2. *Classification/Ordonnancement des termes-clés candidats*

L'étape de classification/ordonnancement intervient après l'extraction des termes-clés candidats. Son rôle est de déterminer quels sont, parmi les candidats, les termes-clés du document analysé. La classification est majoritairement utilisée par les méthodes supervisées. Les méthodes non-supervisées, quant à elles, effectuent en général un ordonnancement des candidats. Dans cet article, nous nous intéressons aux méthodes non-supervisées, nous ne présentons donc que ces dernières. De plus, les différences notables entre différentes méthodes supervisées résident principalement dans le choix du classifieur (classifieur naïf bayésien, arbres de décisions, perceptron multi-couches, etc) ou des traits (TF-IDF, première position, parties du discours, etc.).

Les méthodes non-supervisées d'extraction automatique de termes-clés emploient des techniques très différentes, allant du simple usage de mesures statistiques (Spärck Jones, 1972 ; Paukkeri et Honkela, 2010) au groupement des mots par fréquence de co-occurrences (Liu *et al.*, 2009), en passant par l'utilisation de modèles de langues obtenus à partir de données non-annotées (Tomokiyo et Hurst, 2003), ou encore la construction d'un graphe de co-occurrences (Mihalcea et Tarau, 2004). Puisque la méthode que nous présentons dans cet article est une méthode dite « à base de graphe », nous nous intéressons ici à cette dernière catégorie de méthodes.

Mihalcea et Tarau (2004) proposent une méthode d'ordonnancement d'unités textuelles à partir d'un graphe. Leur méthode, utilisée pour le résumé automatique et l'extraction de termes-clés, s'inspire de la méthode PageRank (Brin et Page, 1998, Google) qui détermine l'importance de pages Web grâce à celles qui s'y réfèrent, et celles auxquelles elles se réfèrent. Le plus une page Web est citée par d'autres, le plus elle est importante, et le plus elle est importante, le plus elle donne d'importance aux pages Web auxquelles elle fait référence. Cette notion de référence entre les pages Web est représentée par un graphe dans lequel les nœuds sont des pages Web et les références sont les liens entre elles. Ensuite, une mesure de centralité, inspirée de la mesure de centralité eigenvector, est appliquée pour ordonner les pages Web par importance. Pour l'extraction de termes-clés avec TextRank, les pages Web sont remplacées par les mots (nom et adjectifs) du document analysé et les liens entre eux symbolisent leur(s) co-occurrence(s) dans une fenêtre de 2 mots. Les mots sont ordonnés par importance et les k meilleurs, les mots clés, servent à la génération des termes-clés. Pour ce faire, les mots-clés sont marqués dans le documents et les plus longues séquences de mots-clés adjacents sont extraits comme termes-clés. Dans cette méthode, la précision de l'ordonnancement dépend de la qualité du graphe qui elle même dépend de la fenêtre de co-occurrences. La définition de cette fenêtre est sujette à une intervention manuelle et peut aussi dépendre des propriétés des documents analysés. Dans nos travaux, nous tentons de nous abstraire de cette fenêtre.

Wan et Xiao (2008) proposent la méthode SingleRank. Celle-ci présente deux améliorations à TextRank. Dans un premier temps, les auteurs pondèrent les liens de co-occurrence par le nombre de co-occurrences calculées avec une fenêtres maintenant définit à 10 mots (par exemple, un mot co-occurrent deux fois avec un autre est relié à celui-ci par un poids de 2). Ce poids est ensuite utilisé pour transférer plus ou moins d'importance lors de l'application de l'algorithme d'ordonnancement de TextRank. Dans un second temps, les termes-clés ne sont plus générés, mais ordonnés à partir de la somme du score d'importance des mots qu'ils contiennent. Bien que la méthode SingleRank donne, dans la majorité des cas, des résultats meilleurs que ceux de TextRank, faire la somme du score d'importance des mots pour ordonner les candidats est une approche maladroite. En effet, cela a pour effet de faire monter dans le classement des candidats qui se recouvrent. Ainsi, dans le document *as_2002_000700ar* de la collection DEFT (voir la section 4), le candidat positif « bio-politique » est classé neuvième, alors que les autres candidats contenant « bio-politique » plus d'autres mots non nécessairement importants occupent les classements 2 à 8. Dans nos travaux nous ordonnons les termes-clés candidats en tenant compte de l'importance du sujet qu'ils représentent puis choisissons un représentant par sujet, nous évitons ainsi le problème rencontré avec SingleRank.

Toujours dans l'optique d'utiliser plus d'informations pour améliorer l'efficacité de l'ordonnancement, Wan et Xiao (2008) étendent SingleRank en utilisant des documents voisins (similaires) du document en cours d'analyse. Leur approche (ExpandRank) consiste à observer les co-occurrences dans les documents similaires afin de renforcer ou ajouter des liens dans le graphe initial. Cependant, en fonction de la similarité entre le document analysé et certains documents voisins, des liens non pertinents

peuvent être ajoutés. Pour y remédier, les auteurs utilisent le score de similarité entre les deux documents comme facteur d'atténuation de l'ajout ou du renforcement de liens. Cette approche donne des résultats au delà de ceux de SingleRank, mais il est important de noter que ses performances sont fortement liées à la présence de documents supplémentaires pertinents.

A l'instar de Wan et Xiao (2008), Tsatsaronis *et al.* (2010) tentent d'améliorer TextRank en modifiant le processus de création des liens entre les nœuds du graphe. Dans leur approche, un lien entre deux mots est créé et pondéré en fonction du lien sémantique de ces derniers selon WordNet (Miller, 1995) ou Wikipedia (Milne et Witten, 2008). Les expériences menées par les auteurs montrent de moins bons résultats que TextRank. Toutefois, en biaisant l'ordonnancement en faveur des mots apparaissant dans le titre du document analysé ou bien en ajoutant le TF-IDF dans le calcul de l'importance des mots, leur méthode est capable de donner de meilleurs résultats que TextRank.

L'usage de sujets dans le processus d'ordonnancement avec TextRank est à l'origine proposé par Liu *et al.* (2010). Reposant sur un modèle LDA (Blei *et al.*, 2003, Latent Dirichlet Allocation), leur méthode effectue des ordonnancements biaisés par les sujets du document, puis fusionne les rangs des mots dans chaque ordonnancement afin d'obtenir un ordonnancement global. Dans notre travail, nous émettons aussi l'hypothèse que le sujet auquel appartient une unité textuelle doit jouer un rôle majeur dans le processus d'ordonnancement. Cependant, nous tentons de nous abstraire de l'usage de documents supplémentaires et n'utilisons donc pas le modèle LDA. De plus, il nous semble plus judicieux d'effectuer un seule ordonnancement, en prenant directement en compte l'appartenance d'une unité textuelle à un sujet particulier.

3. Extraction de termes-clés avec TopicRank

TopicRank est une méthode non-supervisée qui extrait les termes-clés d'un document à partir de sa représentation sous la forme d'un graphe de sujets. Elle se différencie des autres méthodes à base de graphe, car, plutôt que de chercher les mots importants du document, elle cherche ses sujets importants, quels qu'en soient leurs vecteurs (termes-clés candidats). Ce nouveau procédé présente l'intérêt de rassembler des informations complémentaires véhiculées par des candidats différents, mais appartenant tout de même au même sujet. Dans un premier temps, les termes-clés candidats sont groupés par sujets, puis les sujets sont ordonnés et enfin, les candidats les plus représentatifs des sujets les plus importants sont extraits comme termes-clés.

3.1. *Identification des sujets*

Le première étape de l'identification des sujets consiste à trouver les unités textuelles qui les véhiculent. Nous choisisons les termes-clés candidats comme étant ces unités textuelles. En ce qui concerne la méthode d'extraction des termes-clés can-

didats, nous nous appuyons sur les observations de Hulth (2003) qui sont que les termes-clés assignés par des humains sont majoritairement des groupes nominaux. Dans ce sens nous extrayons les plus longues séquences de noms (noms propres inclus) et d'adjectifs, tel qu'effectué par Wan et Xiao (2008) et Hasan et Ng (2010). La section 2 précise que ce type de méthode extrait potentiellement moins de candidats négatifs que d'autres méthodes, ce qui est important pour une identification de qualité des sujets. En addition, nous filtrons les candidats contenant au moins un mots de maximum deux caractères. Ce filtrage est la conséquence de nos observation des documents, de leurs termes-clés associés et des sorties des systèmes d'extraction de termes-clés sans ce filtrage (présence de bruit dû au formatage des documents).

La seconde étape de l'identification des sujets consiste à grouper les termes-clés candidats lorsqu'ils appartiennent au même sujet. Dans le soucis de proposer une méthode ne faisant pas l'usage de données supplémentaires, nous optons pour un groupement quelque peu naïf des candidats. Les candidats sont groupés en fonction d'une similatité de Jaccard (voir l'équation 1) dans laquelle ils sont considérés comme des sacs de mots. En addition, les mots sont tronqués selon la méthode de Porter (1980) afin de considérer identiques ceux qui ont le même radical. Cette mesure est naïve dans le sens où l'ordre des mots, leur ambiguïté et les liens de synonymie ne sont pas pris en compte.

$$\text{sim}(c_1, c_2) = \frac{\|c_1 \cap c_2\|}{\|c_1 \cup c_2\|} \quad [1]$$

Une fois la similarité connue entre tous les candidats deux à deux, nous appliquons l'algorithme de groupement hiérarchique agglomératif (*Hierarchical Agglomerative Clustering – HAC*). Initialement, chaque candidat représente un groupe. À chaque itération de l'algorithme, les deux groupes ayant la plus forte similarité sont unis pour ne former qu'un seul groupe. Afin de ne pas fixer le nombre de sujets à créer comme condition d'arrêt de l'algorithme, nous définissons un seuil de similarité ζ entre les groupes deux à deux. Cette similarité entre deux groupes est calculée en fonction de la similarité entre les candidats de chaque groupe. Il existe trois stratégies pour calculer cette similarité :

– simple : la plus grande valeur de similarité entre les candidats des deux groupes sert de similarité entre eux ;

– complète : la plus petite valeur de similarité entre les candidats des deux groupes sert de similarité entre eux.

– moyenne : la moyenne de toutes les similarités entre les candidats des deux groupes sert de similarité entre eux (compromis entre les stratégies simple et complète) ;

Nous suggérons d'utiliser l'une ou l'autre de ces stratégies en fonction des termes-clés candidats qui sont utilisés. Par exemple, certains ensembles de candidats, tels que les n-grammes pour $n \in 1..m$, contenant de nombreux candidats se recouvrant partiellement risquent de donner lieu à des groupes non consistants avec la stratégie simple. En

revanche, la stratégie complète a tendance à moins regrouper, elle est donc plus adaptée à ces ensembles de candidats. Dans le cas de TopicRank, les termes-clés candidats étant les plus longues séquences de noms et d'adjectifs, la stratégie complète n'est cette fois-ci plus la plus pertinente. Par défaut, nous utilisons la stratégie moyenne.

3.2. *Ordonnancement des sujets*

L'ordonnancement des sujets a pour objectif de trouver quels sont ceux qui ont le plus d'importance dans le document analysé. À l'instar de Mihalcea et Tarau (2004), l'importance des sujets est déterminée à partir d'un graphe.

Les sujets du document analysé composent les nœuds V du graphe complet $G = (V, E)$, E étant l'ensemble des liens entre les nœuds[1]. Le graphe utilisé étant un graphe complet, la pondération de ses arêtes est l'étape la plus importante pour rendre possible un ordonnancement efficace des sujets. Pour celle-ci, nous choisissons d'utiliser la force du liens sémantique entre les sujets. Contrairement à ce qui est fait dans les autres travaux (Wan et Xiao, 2008 ; Tsatsaronis *et al.*, 2010 ; Liu *et al.*, 2010), nous ne représentons pas cette force avec le nombre de co-occurrences calculées dans une fenêtre de mots définie manuellement, mais nous utilisons la distance, dans le document, entre les mots des sujets :

$$\text{poids}(s_i, s_j) = \sum_{c_i \in s_i} \sum_{c_j \in s_j} \text{dist}(c_i, c_j) \quad [2]$$

$$\text{dist}(c_i, c_j) = \sum_{p_i \in \text{pos}(c_i)} \sum_{p_j \in \text{pos}(c_j)} \frac{1}{|p_i - p_j|} \quad [3]$$

où $\text{poids}(s_i, s_j)$ est le poids de l'arête entre les sujets s_i et s_j, et où $\text{dist}(c_i, c_j)$ représente la force sémantique entre les candidats c_i et c_j, calculée à partir de leurs positions respectives, $\text{pos}(c_i)$ et $\text{pos}(c_j)$, dans le document.

Une fois la construction du graphe, l'algorithme d'ordonnancement de TextRank est utilisé. Celui-ci se fonde sur le principe de « vote », c'est à dire qu'un sujet fortement connecté à un autre sujet est fortement recommandé par le dernier, il gagne donc de l'importance. De ce fait, un sujet connecté à un autre sujet très important gagne aussi plus d'importance :

$$\text{importance}(s_i) = (1 - \lambda) + \lambda \times \sum_{s_j \in V_i} \frac{\text{poids}(s_i, s_j) \times \text{importance}(s_j)}{\sum_{s_k \in V_j} \text{poids}(s_j, s_k)} \quad [4]$$

où V_i est l'ensemble des sujets connectés au sujet s_i[2] et où λ est un facteur d'atténuation définit par défaut à 0,85 par Brin et Page (1998).

1. $E = \{(v_1, v_2) \mid \forall v_1, v_2 \in V, v_1 \neq v_2\}$, car G est un graphe complet.
2. $V_i = \{v_i \mid \forall v_j in V, v_j \neq v_i\}$, car G est un graphe complet.

3.3. *Sélection des termes-clés*

La sélection des termes-clés est la dernière étape de TopicRank. Elle consiste à choisir le terme-clé candidat le plus représentatif d'un sujet. Pour chacun des k sujets les plus importants, ce principe de sélection donne lieu à k termes-clés non redondant et couvrant exactement k sujets.

La difficulté de ce principe de sélection réside dans la capacité à trouver parmi plusieurs termes-clés candidats d'un même sujet celui qui le représente le mieux. Nous distinguons trois stratégies de sélection pouvant répondre à ce problème :

– la première position : en supposant qu'un sujet est tout d'abord introduit dans sa forme la plus appropriée, le terme-clé candidat sélectionné pour un sujet est celui qui apparaît en premier dans le document analysé ;

– la fréquence : en supposant que la forme la plus représentative d'un sujet est sa forme la plus fréquente, le terme-clé candidat sélectionné pour un sujet est celui qui est le plus fréquent dans le document analysé ;

– le centroïde : le terme-clé candidat sélectionné pour un sujet est celui qui est le plus similaire aux autres (voir l'équation 1).

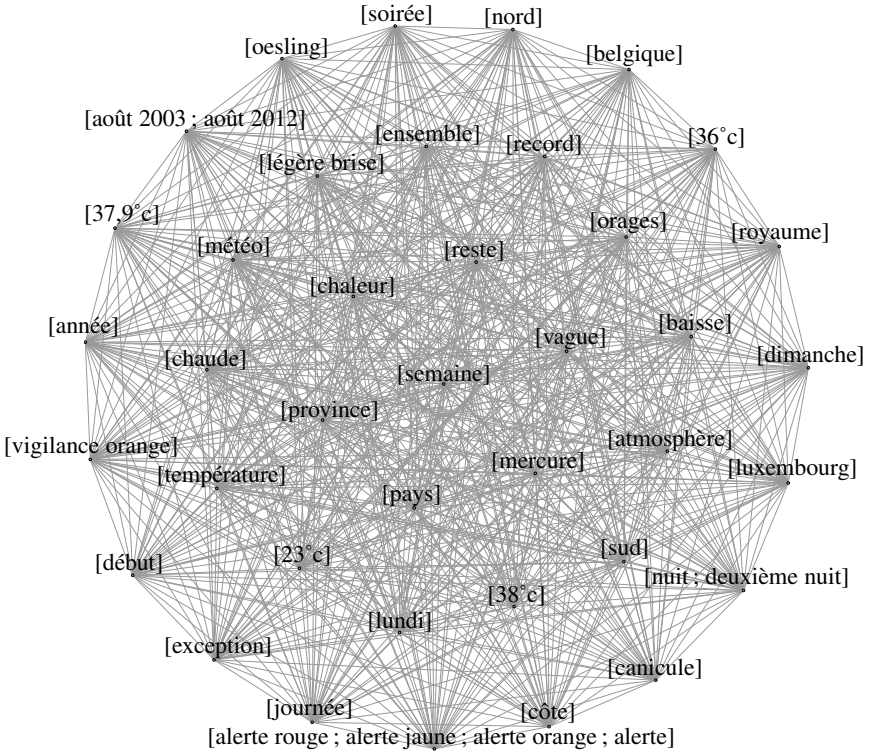
Parmi ces trois stratégies, celle qui semble la plus appropriée est la stratégie qui se fonde sur la première position des termes-clés candidats. En effet, sélectionner les candidats les plus fréquents risque de favoriser l'extraction de formes abrégées ou de concepts inhérents. Par exemple, dans la collection SemEval (voir la section 4), le document *C-17* parle de « réseaux à commutation de paquets » (*packet-switched networks*), mais le candidat le plus fréquent dans le sujet correspondant est le concept inhérent « réseau » (*network*). Extraire le centroïde de chaque groupe risque d'avoir un effet semblable, car « réseau » est le sous-composant de nombreux autres candidats du sujet : « réseau étendu » (*wide area network*), « réseaux locaux » (*local area networks*),- « réseaux informatisés de communication » (*computer-communication networks*), etc.

La figure 2 donne un exemple d'extraction de termes-clés, à partir d'un article de journal, avec TopicRank. Nous observons un groupement correct de toutes les variantes d'« alertes », mais aussi un groupement erroné de « août 2003 » avec « août 2012 ». Dans ce dernier cas, TopicRank est tout de même capable d'extraire « août 2012 », grâce à la sélection du candidat apparaissant en premier. Globalement, l'extraction des termes-clés est correcte et huit termes-clés sur les dix extraits ont aussi été donnés par des humains.

4. Évaluation

Pour valider notre approche, nous réalisons une série d'évaluations visant à déterminer la configuration optimale de TopicRank, le comparer aux autres méthodes existantes et quantifier chacun de ses apports.

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 extraits par des humains :
luxembourg ; alerte ; météo ; belgique ; août 2012 ; chaleur ; température ; chaude ; canicule ; orange ; la plus chaude
Termes-clés extraits par TopicRank :
luxembourg ; alerte ; nuit ; belgique ; août 2012 ; chaleur ; température ; chaude ; canicule ; dimanche

Figure 2. *Extraction des termes-clés du document 44960 de WikiNews (voir la section 4), avec TopicRank.*

4.1. *Environnement expérimental*

4.1.1. *Données de test*

Les collections de données présentées ci-dessous sont utilisées lors de toutes les évaluations. Afin de suivre Hasan et Ng (2010) qui soulignent l'importance d'évaluer une méthode avec des collections de données aux configurations différentes pour mieux observer et comprendre son comportement, les collections de données utilisées sont différentes en termes de langue, nature, taille des documents et types d'annotateur (auteurs, lecteurs ou les deux).

DUC (Over, 2001) est une collection (en anglais) issue des données de la campagne d'évaluation DUC-2001. Cette campagne d'évaluation concerne les méthodes de résumé automatique, elle ne contient donc originellement pas d'annotations en termes-clés. Cependant, les 308 articles journalistiques de la partie test de DUC-2001 sont annotés par Wan et Xiao (2008). Lors de nos expériences, nous utilisons ces 308 documents.

SemEval (Kim *et al.*, 2010) est la collection fournie lors de la campagne d'évaluation SemEval-2010 pour la tâche d'extraction automatique de termes-clés. Cette collection contient 284 articles scientifiques (conférences et ateliers) issus de la librairie numérique ACM (en anglais). La collection est répartie en trois sous-ensembles, un ensemble de 40 documents d'essais, un ensemble de 144 documents d'entraînement et un ensemble de 100 documents de test. Lors de nos expériences, nous utilisons les 100 documents de l'ensemble de test. En ce qui concerne les termes-clés associés aux documents, ils sont donnés par les auteurs et des lecteurs.

WikiNews[3] est une collection de 100 articles journalistiques français extraits à partir du site Web WikiNews[4] entre les mois de mai et décembre 2012. Chaque document est annoté par au moins trois étudiants, les termes-clés des différents étudiants sont groupés et les redondances lexicales sont automatiquement supprimées.

DEFT (Paroubek *et al.*, 2012) est la collection fournie lors de la campagne d'évaluation DEFT-2012 pour la tâche d'extraction automatique de termes-clés. Celle-ci contient 234 documents français issus de quatre revues de Sciences Humaines et Sociales. La collection est divisée en deux sous-ensembles, un ensemble d'entraînement contenant 141 documents et un ensemble de test contenant 93 documents. Lors de nos expériences, nous utilisons les 93 documents de l'ensemble de test. Seuls les termes-clés d'auteurs sont disponibles pour cette collection.

Le tableau 1 donne les statistiques extraites des quatre collections de données présentées ci-dessus. Les données sont divisées en deux langues (anglais et français), avec pour chaque langues une collections de documents courts (articles journalistiques) et une collection de documents de plus grande taille (articles scientifiques). Il est aussi important de noter qu'en fonction du type d'annotateurs, le nombre de termes-clés

3. `https://github.com/adrien-bougouin/WikinewsKeyphraseCorpus`
4. `http://fr.wikinews.org/`

Statistique	DUC	SemEval	WikiNews	DEFT
Langue	Anglais	Anglais	Français	Français
Nature	Journalistique	Scientifique	Journalistique	Scientifique
Annotateurs	Lecteurs	Auteurs & Lecteurs	Lecteurs	Auteurs
Documents	308	100	100	93
Mos/document	900,7	5177,7	308,5	6839,4
Termes-clés/document	8,1	14,7	9,6	5,2
Mots/termes-clés	2,1	2,1	1,7	1,6
Termes-clés manquants	3,5%	22,1%	7,6%	21,1%

Tableau 1. *Statistiques sur les données de test utilisées. En accord avec l'évaluation effectuée lors de nos expériences, la proportion de termes-clés manquant est déterminée sans tenir compte de la flexions des mots.*

associés varie, de même que le nombre de termes-clés n'apparaissant pas dans les documents.

4.1.2. *Prétraitement*

Chaque document des collections de données utilisées subissent les mêmes prétraitements. Ils sont tout d'abord segmentés en phrases, puis en mots et enfin étiquetés en parties du discours. La segmentation en mots est effectuée par le TreeBankWordTokenizer disponible avec la librairie python NLTK (Bird *et al.*, 2009, *Natural Language ToolKit*), pour l'anglais, et par l'outil Bonsai du Bonsai PCFG-LA parser[5], pour le français. Quant à l'étiquetage en parties du discours, il est réalisé avec le Stanford POS tagger (Toutanova *et al.*, 2003), pour l'anglais, et avec MElt (Denis et Sagot, 2009), pour le français.

4.1.3. *Mesures d'évaluation*

Les performances des méthodes d'extraction de termes-clés sont exprimées en termes de précision (P), rappel (R) et f-score (f1-mesure, F). Afin de réduire le problème de termes-clés manquants, les comparaisons sont effectuées avec la forme non flexionnelle des mots des termes-clés extraits et des termes-clés de référence.

4.1.4. *Méthodes de référence pour l'extraction de termes-clés*

Dans nos expérimentations, nous comparons TopicRank avec trois autres méthodes non-supervisées d'extraction automatique de termes-clés. Nous choisissons TextRank et SingleRank, les deux méthodes qui sont la fondation des méthodes à base de graphe, et la pondération TF-IDF. Cette dernière consiste à donner un score aux termes-clés

5. `http://alpage.inria.fr/statgram/frdep/fr_stat_dep_parsing.html`

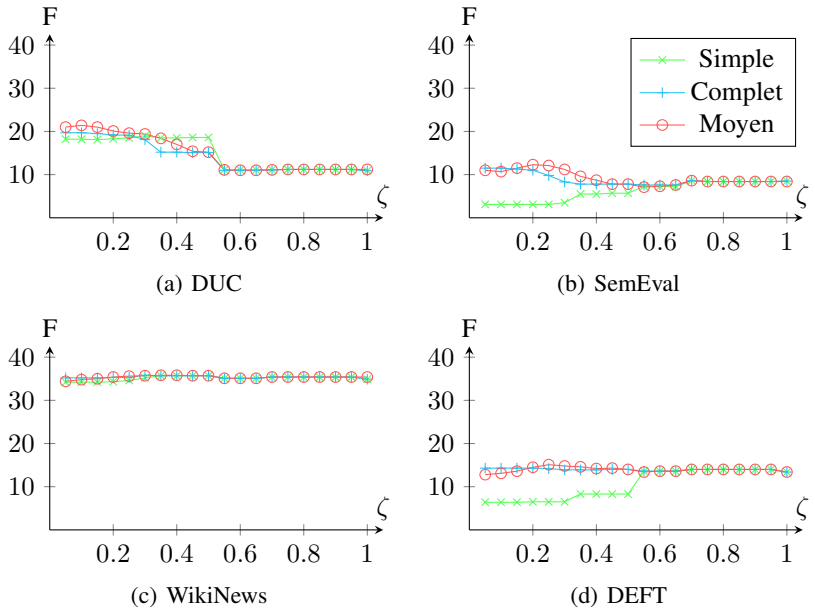
candidats en faisant la somme des poids TF-IDF des mots qui les composent, puis à sélectionner ceux ayant le plus haut score.

Toutes les méthodes sont réimplémentées. lorsque celles-ci ont des troncs commun avec TopicRank, elles bénéficies des même composants. Afin d'améliorer le résultat des méthodes de référence, leurs sorties sont filtrées pour de supprimer les termes-clés redondant. Conformément au processus d'évaluation mis en œuvre, deux termes-clés sont considérés redondants lorsqu'ils ont le même radical. Cette suppression ne dégrade en rien les résultats et a pour effet d'extraire des termes-clés qui ne sont pas parmi les 10 premiers avant le filtrage.

4.2. *Configuration empirique de TopicRank*

À ce stade des expérimentations, nous tentons de déterminer quels sont les paramètres optimaux pour TopicRank. En effet, TopicRank possède trois points de variabilités : le seuil de groupement (ζ), la stratégie de groupement (simple, complète ou moyenne) et la stratégie de sélection du terme-clé candidat le plus représentatif d'un sujet. Deux expériences sont réalisées, l'une pour déterminer le groupement optimale (variation du seuil ζ et de la stratégie de groupement) et l'autre pour déterminer la sélection des termes-clés la plus optimale.

La figure 3 présente les résultats de TopicRank lorsque nous faisons varier le seuil ζ avec un pas de 0,05 pour toutes les stratégies de regroupement. La stratégie de sélection d'un termes-clés par sujet utilisée est celle qui consiste à sélectionner le candidat qui apparaît en premier dans le document, pour chaque sujet. Globalement, chaque stratégie de groupement a un comportement qui lui est propre, jusqu'à un certain point de convergence lorsque ζ vaut 0,55. Avec la stratégie simple, les résultats s'améliorent lorsque le seuil ζ augmente. Du fait qu'elle ne prend en compte que la similarité maximale entre deux candidats de deux groupes, cette stratégie à tendance à trop grouper et donc à créer des groupes contenant parfois plusieurs sujets. L'augmentation du seuil ζ a pour effet de restreindre cette tendance et la qualité du groupement s'améliore. En opposition, la stratégie complète, qui a le fonctionnement inverse, voie ses résultats se dégrader lorsque ζ augmente. Finalement, la stratégie moyenne, qui agit en tant que compromis, semble moins sensible aux variations de ζ, même si nous observons une dégradation des résultats jusqu'au point de convergence. Ce point de convergence correspond au moment où les groupes sont majoritairement composés de variantes flexionnelles où de candidats dont les plus long inclus les autres (par exemple, le sujet qui ne contient que « nouvelles églises », « nouvelles églises indépendantes protestantes », « nouvelle église indépendante » et « nouvelles églises indépendantes », avec le groupement moyen et $\zeta = 0{,}55$, contient aussi « églises évangéliques », ainsi que d'autres, lorsque ζ est plus faible). Après observation des résultats de cette expérience, le seuil ζ est fixé à 0,25 pour le reste des expériences. De même, la stratégie de groupement utilisée dans la suite est la stratégie moyenne. En effet, la stratégie moyenne est celle qui donne les meilleurs résultats et ceux-ci sont obtenus lorsque ζ varie entre



(a) DUC (b) SemEval

(c) WikiNews (d) DEFT

Figure 3. *Résultats de l'extraction de 10 termes-clés, avec TopicRank, en fonction de la stratégie de regroupement et de la valeur du seuil de similarité ζ.*

0,20 et 0,35 suivant les collections, 0,25 étant la valeur fournissant le meilleur compromis pour chacune d'elles.

La figure 4 présente les résultats obtenus avec TopicRank et les différentes stratégies de sélection d'un terme-clé candidat par sujet. Ceux-ci confirment ce qui est dit dans la section 3, concernant le bien fondé de la sélection des candidats les plus fréquents ou des centroïdes. Ces dernières stratégies ont tendance à sélectionner des concepts inhérents qui jouent un rôle crucial lors du groupement, mais qui ne sont pas les candidats les plus représentatifs des sujets. Bien que la sélection à partir de la première position des candidats donne des résultats satisfaisant, nous remarquons qu'il existe encore une marge de progression importante. En effet, les valeurs indiquées par la borne haute représentent les résultats qui pourraient être obtenue avec une stratégie toujours capable de sélectionner un candidat positif. Cette marge de progression allant de 4,2 à 19,0 points de f-score est encourageante pour de futurs travaux.

4.3. *Comparaison de TopicRank avec l'existant*

Le tableau 2 montre les performances de TopicRank comparées à celles des trois systèmes de référence. Globalement, TopicRank donne de meilleurs résultats que les méthodes de référence utilisées. Comparée à la méthode TF-IDF, TopicRank donne

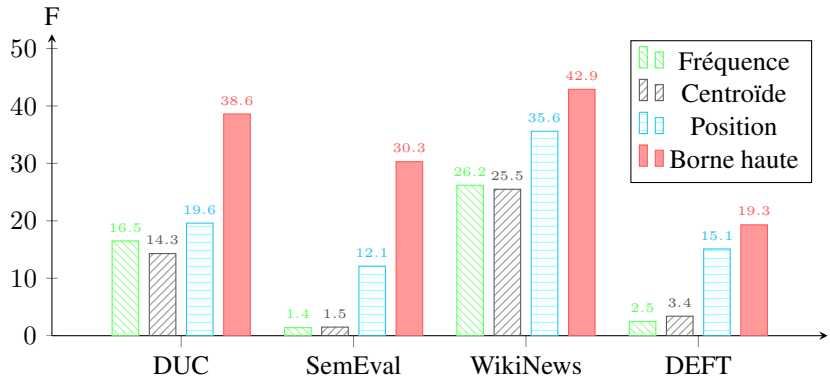


Figure 4. *Résultats de l'extraction de 10 termes-clés, avec TopicRank, en fonction des différentes sélections de termes-clés candidats par sujet.*

Méthode	DUC			SemEval			WikiNews			DEFT		
	P	R	F	P	R	F	P	R	F	P	R	F
TF-IDF	23,8	30,7	26,4	13,2	8,9	10,5	33,9	35,9	34,3	10,3	19,1	13,2
TextRank	4,9	5,4	5,0	7,9	4,5	5,6	9,3	8,3	8,6	4,9	7,1	5,7
SingleRank	22,3	28,4	24,6	4,6	3,2	3,7	19,4	20,7	19,7	4,5	9,0	5,9
TopicRank	17,7	22,6	19,6	14,9	10,3	**12,1**†	35,0	37,5	**35,6**†	11,7	21,7	**15,1**†

Tableau 2. *Résultats de l'extraction de 10 termes-clés avec TF-IDF, TextRank, SingleRank et TopicRank.* † *indique une amélioration significative de TopicRank vis-à-vis de TextRank et SingleRank, à 0,001 pour le t-test de Student.*

de meilleurs résultats pour SemEval, WikiNews et DEFT. Cette supériorité vis-à-vis de TF-IDF est importante à noter, car cette méthode obtiens de bons résultats grâce à des statistiques extraites de documents supplémentaires (apprentissage non supervisé), alors que TopicRank n'utilise que le document à analysé. Comparée au autre méthodes à base de graphe, TopicRank donne des résultats, cette fois-ci, significativement meilleurs. Ceci confirme donc que le groupement des candidats contribue à rassembler des informations utiles lors de l'ordonnancement. En ce qui concerne DUC, TopicRank est toujours significativement meilleure que TextRank, mais pas vis-à-vis des autres méthodes. Une des raison pour lesquels les résultats sont moins satisfaisant pour DUC est que la stratégie de sélection des candidats les plus représentatifs des sujets n'est pas adaptée (voir la borne haute de la figure 4). Une analyse plus approfondie des différents apports de TopicRank peut aussi donner une piste sur les raisons de ces moins bons résultats.

Méthode	DUC			SemEval			WikiNews			DEFT		
	P	R	F	P	R	F	P	R	F	P	R	F
SingleRank	22,3	28,4	**24,6**	4,6	3,2	3,7	19,4	20,7	19,7	4,5	9,0	5,9
+complet	22,2	28,1	24,5	5,5	3,8	4,4	20,0	21,4	20,3	4,4	9,0	5,8
+candidats	10,0	13,1	11,2	9,6	7,0	8,0†	28,6	30,1	28,9†	10,5	19,7	13,5†
+sujets	18,4	23,6	20,5	14,7	10,2	11,9†	31,0	32,8	31,4†	11,5	21,4	14,8†
TopicRank	17,7	22,6	19,6	14,9	10,3	**12,1**†	35,0	37,5	**35,6**†	11,7	21,7	**15,1**†

Tableau 3. *Résultats de l'extraction de 10 termes-clés avec chacune des contributions de TopicRank appliquées séparément à SingleRank. † indique une amélioration significative vis-à-vis de SingleRank, à 0,001 pour le t-test de Student.*

Dans le but de confirmer la pertinence de tous les apports de TopicRank, nous réalisons une expérience supplémentaire dans laquelle la méthode SingleRank est modifiée de sorte qu'elle ordonne les mots avec un graphe complet, qu'elle ordonne les termes-clés candidats à la place des mots ou qu'elle ordonne les sujets. Ces trois variantes de SingleRank sont présentées dans le tableau 3. Globalement, l'usage des termes-clés candidats, ou des sujets, induit une amélioration significative des performances de SingleRank, avec une amélioration plus importante en utilisant les sujets. L'usage d'un graphe complet, quant à lui, n'améliore pas significativement les résultats de SingleRank. Ceux-ci sont compétitifs avec ceux obtenus en construisant un graphe de co-occurrences. Nous pensons cependant que l'usage de ce graphe complet est à privilégier afin d'éviter cette fenêtre qui doit être définit manuellement. En ce qui concerne la collection DUC, nous observons une perte de performances induites lors de la construction du graphe à partir des termes-clés candidats. Cette perte de performance s'explique par le fait qu'il y a peu de répétition des candidats dans les documents de DUC et donc moins d'informations pour la création de liens. Nous l'observons aussi dans la figure 4, où l'extraction des candidats les plus fréquents par sujet est compétitive avec l'extraction des candidats apparaissant en premier dans le document. TODO trouver un exemple

4.4. *Analyse des sujets détectés*

Dans cette section, nous analysons les groupes créés et tentons de déterminer quelles sont les causes des groupements incorrects.

À partir des groupements obtenus pour DEFT, nous observons trois causes de groupements erronés, dont deux majeures. Dans un premier temps, nous remarquons que les adjectifs ayant attraits au contexte général du document sont la principale cause d'erreurs de groupement. En effet, ces adjectifs, tels que « européen » ou « économique », sont présents dans de nombreux candidats et ont donc tendance à tous les rapprocher et ainsi favoriser leur groupement. Par exemple, dans le docu-

ment *as_2002_000707ar* qui examine l'organisation du dialogue politique entre les membres de la commission européenne, les termes-clés candidats « imaginaire européen », « rhétorique européenne », « chose européenne » et « culture européenne » sont groupés à cause de l'adjectif « européen ». Dans un second temps, nous observons que deux candidats non reliés contenant deux mots (candidats les plus fréquent après les candidats contenant un seul mot) et un mot en commun sont plus difficilement mis dans des sujets différents. Ainsi, toujours dans le document as_2002_000707ar, des candidats tels que « dimension spatiale » et « dimension culturelle » sont considérés comme appartenant au même sujets. En addition à ces deux causes principales d'erreurs de détection de sujets, l'utilisation des radicaux lors du calcul de la similarité de Jaccard induit elle aussi quelques erreurs. Dans le document *as_2002_000702ar*, l'algorithme de groupement à, par exemple, groupé les deux candidats « empire ottoman » et « définition empirique », « empire » et « empirique » ayant tous les deux le radical « empir » selon la méthode de Porter. Pour résoudre ce problème, l'usage des lemmes à la place des radicaux est une alternative à envisager dans de futurs travaux.

5. Conclusion et perspectives

Dans ce travail, nous proposons une méthode à base de graphe pour l'extraction non-supervisée de termes-clés. Cette méthode groupe les termes-clés candidats par sujets, détermine quels sont ceux les plus importants, puis extrait le terme-clé candidat qui représente le mieux chacun des sujets les plus importants. Cette nouvelle méthode offre plusieurs avantages vis-à-vis des précédentes à base de graphe. Dans un premier temps, le groupement des termes-clés potentiels en sujets distincts permet le rassemblement d'informations auparavant éparpillées et dans un second temps, le choix d'un seul terme-clé pour représenter l'un des sujets les plus importants permet d'extraire un ensemble de termes-clés non redondant – pour k termes-clés extraits, exactement k sujets sont couverts. Finalement, le graphe est désormais complet et ne requière plus de fenêtre de co-occurrences définie manuellement.

Les bons résultats de notre méthodes montrent la pertinence d'un groupement en en sujets des candidats pour ensuite les ordonner. Les expériences supplémentaires montrent aussi que la stratégie de sélection du terme-clé candidat le plus représentatif d'un sujet joue un rôle crucial. La stratégie actuellement utilisée pourrait ainsi être améliorée de sorte que les résultats soient significativement améliorer (pour un gain maximum allant de 4,2 à 22,1 points de f-score).

Dans de futurs travaux, il est envisagé d'améliorer le groupement en sujets et la sélection du terme-clé candidat le plus représentatif pour chacun d'eux. Ces deux points sont cruciaux et nécessitent un travail plus approfondit linguistiquement.

Le groupement actuellement effectué est un groupement naïf qui ne prend en compte ni l'ambiguïté d'un mots, ni la relation de synonymie entre deux mots. L'ajout de connaissances concernant les synonymes permettrait de créer des sujets plus consistants et la désambiguïsation éviterait un groupement systématique des termes-clés can-

didats ayant un ou plusieurs mots en commun. D'un point de vu plus technique, il est aussi envisagé d'explorer différentes techniques de groupement, dont le groupement spectral (*spectral clustering*) qui, dans d'autres travaux portant sur l'extraction automatique de termes-clés (Liu *et al.*, 2009), montre de meilleures performances que le groupement hiérarchique agglomératif.

En ce qui concerne la stratégie de sélection des termes-clés candidats les plus représentatifs des sujets, une étude détaillées des caractéristiques des termes-clés pourrait orienter notre travail vers des critères plus efficaces que la première position des candidats dans le document. Un apprentissage supervisé à partir de certains critère est aussi envisageable, au même titre que l'usage de méthodes d'optimisation telles que celle utilisée par Ding *et al.* (2011) dans leur méthode d'extraction automatique de termes-clés.

Remerciements

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

6. Bibliographie

Bird S., Klein E., Loper E., *Natural Language Processing with Python*, O'Reilly Media, 2009.

Blei D. M., Ng A. Y., Jordan M. I., « Latent Dirichlet Allocation », *Journal of Machine Learning Research*, vol. 3, p. 993-1022, 2003.

Brin S., Page L., « The Anatomy of a Large-Scale Hypertextual Web Search Engine », *Computer Networks and ISDN Systems*, vol. 30, n° 1, p. 107-117, 1998.

Denis P., Sagot B., « Coupling an Annotated Corpus and a Morphosyntactic Lexicon for State-of-the-Art POS Tagging with Less Human Effort », *Proceedings of the 23rd Pacific Asia Conference on Language, Information and Computation (PACLIC)*, City University of Hong Kong, Hong Kong, p. 110-119, December, 2009.

Ding Z., Zhang Q., Huang X., « Keyphrase Extraction from Online News Using Binary Integer Programming », *Proceedings of the 5th International Joint Conference on Natural Language Processing*, Asian Federation of Natural Language Processing, Chiang Mai, Thailand, p. 165-173, November, 2011.

Eichler K., Neumann G., « DFKI KeyWE : Ranking Keyphrases Extracted from Scientific Articles », *Proceedings of the 5th International Workshop on Semantic Evaluation*, Association for Computational Linguistics, Stroudsburg, PA, USA, p. 150-153, 2010.

Han J., Kim T., Choi J., « Web Document Clustering by Using Automatic Keyphrase Extraction », *Proceedings of the 2007 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology*, IEEE Computer Society, Washington, DC, USA, p. 56-59, 2007.

Hasan K. S., Ng V., « Conundrums in Unsupervised Keyphrase Extraction : Making Sense of the State-of-the-Art », *Proceedings of the 23rd International Conference on Computational*

Linguistics : Posters, Association for Computational Linguistics, Stroudsburg, PA, USA, p. 365-373, 2010.

Hulth A., « Improved Automatic Keyword Extraction Given More Linguistic Knowledge », *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, Stroudsburg, PA, USA, p. 216-223, 2003.

Kim S. N., Kan M.-Y., Baldwin T., « An Unsupervised Approach to Domain-Specific Term Extraction », *Proceedings of the 2009 Australasian Language Technology Association Workshop*, 2009.

Kim S. N., Medelyan O., Kan M.-Y., Baldwin T., « SemEval-2010 task 5 : Automatic Keyphrase Extraction from Scientific Articles », *Proceedings of the 5th International Workshop on Semantic Evaluation*, Association for Computational Linguistics, Stroudsburg, PA, USA, p. 21-26, 2010.

Litvak M., Last M., « Graph-Based Keyword Extraction for Single-Document Summarization », *Proceedings of the Workshop on Multi-Source Multilingual Information Extraction and Summarization*, Association for Computational Linguistics, Stroudsburg, PA, USA, p. 17-24, 2008.

Liu Z., Huang W., Zheng Y., Sun M., « Automatic Keyphrase Extraction Via Topic Decomposition », *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, Stroudsburg, PA, USA, p. 366-376, 2010.

Liu Z., Li P., Zheng Y., Sun M., « Clustering to Find Exemplar Terms for Keyphrase Extraction », *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing : Volume 1*, Association for Computational Linguistics, Stroudsburg, PA, USA, p. 257-266, 2009.

Medelyan O., Witten I. H., « Domain-Independent Automatic Keyphrase Indexing with Small Training Sets », *Journal of the American Society for Information Science and Technology*, vol. 59, no 7, p. 1026-1040, may, 2008.

Mihalcea R., Tarau P., « TextRank : Bringing Order Into Texts », *in* Dekang Lin, Dekai Wu (eds), *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, Barcelona, Spain, p. 404-411, July, 2004.

Miller G. A., « WordNet : a Lexical Database for English », *Communications of the Association for Computational Linguistics*, vol. 38, no 11, p. 39-41, 1995.

Milne D., Witten I. H., « An Effective, Low-Cost Measure of Semantic Relatedness Obtained from Wikipedia Links », *Proceeding of Association for the Advancement of Artificial Intelligence Workshop on Wikipedia and Artificial Intelligence : an Evolving Synergy*, p. 25-30, 2008.

Over P., « Introduction to DUC-2001 : an Intrinsic Evaluation of Generic News Text Summarization Systems », *Proceedings of DUC 2001 Document Understanding Conference*, 2001.

Paroubek P., Zweigenbaum P., Forest D., Grouin C., « Indexation libre et contrôlée d'articles scientifiques. Présentation et résultats du défi fouille de textes DEFT2012 (Controlled and Free Indexing of Scientific Papers. Presentation and Results of the DEFT2012 Text-Mining Challenge) [in French] », *Proceedings of the Joint Conference JEP-TALN-RECITAL 2012, Workshop DEFT 2012 : DÉfi Fouille de Textes (DEFT 2012 Workshop : Text Mining Challenge)*, ATALA/AFCP, Grenoble, France, p. 1-13, June, 2012.

Paukkeri M.-S., Honkela T., « Likey : Unsupervised Language-Independent Keyphrase Extraction », *Proceedings of the 5th International Workshop on Semantic Evaluation*, Association for Computational Linguistics, Stroudsburg, PA, USA, p. 162-165, 2010.

Porter M. F., « An Algorithm for Suffix Stripping », *Program : Electronic Library and Information Systems*, vol. 14, n° 3, p. 130-137, 1980.

Spärck Jones K., « A Statistical Interpretation of Term Specificity and its Application in Retrieval », *Journal of Documentation*, vol. 28, n° 1, p. 11-21, 1972.

Tomokiyo T., Hurst M., « A Language Model Approach to Keyphrase Extraction », *Proceedings of the ACL 2003 Workshop on Multiword Expressions : Analysis, Acquisition and Treatment - Volume 18*, Association for Computational Linguistics, Stroudsburg, PA, USA, p. 33-40, 2003.

Toutanova K., Klein D., Manning C. D., Singer Y., « Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network », *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*, Association for Computational Linguistics, Stroudsburg, PA, USA, p. 173-180, 2003.

Tsatsaronis G., Varlamis I., Nørvåg K., « SemanticRank : Ranking Keywords and Sentences Using Semantic Graphs », *Proceedings of the 23rd International Conference on Computational Linguistics*, Association for Computational Linguistics, Stroudsburg, PA, USA, p. 1074-1082, 2010.

Turney P. D., « Learning Algorithms for Keyphrase Extraction », *Information Retrieval*, vol. 2, n° 4, p. 303-336, may, 2000.

Wan X., Xiao J., « Single Document Keyphrase Extraction Using Neighborhood Knowledge », *Proceedings of the 23rd National Conference on Artificial Intelligence - Volume 2*, AAAI Press, p. 855-860, 2008.

Witten I. H., Paynter G. W., Frank E., Gutwin C., Nevill Manning C. G., « KEA : Practical Automatic Keyphrase Extraction », *Proceedings of the 4th ACM Conference on Digital Libraries*, ACM, New York, NY, USA, p. 254-255, 1999.