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 les expressions polylexicales qui représentent le contenu principal d'un document. Ils sont utiles pour diverses applications telles que l'indexation automatique ou le résumé automatique, mais ne sont cependant pas disponibles pour la plupart des documents. La quantité de ces documents étant de plus en plus importante, l'extraction manuelle des termes-clés n'est pas envisageable et la tâche d'extraction automatique de termes-clés suscite alors l'intérêt des chercheurs. Dans cet article nous présentons TopicRank, une méthode non-supervisée à base de graphe pour l'extraction de termes-clés. Cette méthode groupe les termes-clés candidats par sujet, ordonne les groupes obtenus, et extrait des meilleurs groupes le terme-clé candidat le plus représentatif. 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 fo 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

Un terme-clé est un mot ou une expression polylexicale permettant de caractériser le contenu d'un document. Un ensemble de termes-clés permet ainsi de définir les principales thématiques abordées dans un document. Du fait de cette propriété synthétique, les ensembles de termes-clés sont utilisés dans de nombreuses applications du Traitement Automatique des Langues (TAL) telles que l'indexation automatique de documents (Medelyan et Witten, 2008), le résumé automatique (D'Avanzo et Magnini, 2005), la compression multi-phrase (Boudin et Morin, 2013) ou encore la classification de documents (Han *et al.*, 2007). Pourtant, de nombreux documents, tels que ceux accessibles par Internet, ou collections de documents, telles que les actes de conférences, n'en sont pas accompagnées. La quantité de document à traiter est aujourd'hui trop importante pour que l'annotation de leurs termes-clés soit effectuée manuellement. C'est pourquoi de nombreux chercheurs se penchent sur la problématique de l'extraction automatique de termes-clés.

L'extraction automatique de termes-clés consiste à sélectionner dans un document les unités textuelles les plus importantes, c'est-à-dire celles qui représentent le contenu principale du document. Bien que des documents supplémentaires peuvent être utilisés, l'extraction de termes-clés ne concerne qu'un seul document à la fois. Parmi les différentes méthodes d'extraction automatique de termes-clés proposées dans la littérature, deux grandes catégories émergent : les méthodes supervisées et les méthodes non-supervisées. Les premières réduisent la tâche d'extraction de termes-clés en une tâche de classification binaire (Witten *et al.*, 1999), où il s'agit d'attribuer la classe « *terme-clé* » ou « *non terme-clé* » aux différents candidats extraits à partir du document. Une collection de documents annotés en termes-clés est alors nécessaire pour l'apprentissage du modèle de classification. Les méthodes non-supervisées, quant à elles, attribuent un score d'importance à chaque candidat en fonction de divers indicateurs comme par exemple la fréquence, le nombre de co-occurrences ou la position dans le document. Bien que les méthodes supervisées soient en général plus performantes, la faible quantité de documents annotés en termes-clés disponibles couplée à la forte dépendance des modèles de classification vis-à-vis du type de documents sur lesquels ils ont été appris, poussent les chercheurs à s'intéresser de plus en plus aux méthodes non-supervisées.

Les méthodes d'extraction de termes-clés non-supervisées les plus étudiées sont sans conteste celles basée sur TextRank (Mihalcea et Tarau, 2004), qui est une méthode d'ordonnancement d'unités textuelles à partir de graphe. Les graphes sont un moyen naturel de représenter les unités textuelles et les relations qui les relient, et ils sont utilisés dans de nombreuses applications du TAL (Kozareva *et al.*, 2013). Pour l'extraction de termes-clés, l'idée est de représenter le document sous la forme d'un graphe dans lequel les nœuds correspondent aux mots et les arêtes correspondent à des relations de co-occurrences dans une fenêtre de mots. Un score d'importance est alors calculé pour chaque mot selon un principe de recommandation, c'est-à-dire un mot est d'autant plus important s'il co-occurre avec un grand nombre de mots et si les mots

avec lesquels il co-occurre sont eux aussi importants. Les mots les plus importants sont ensuite assemblés pour générer les termes-clés.

Dans cet article, nous présentons TopicRank, une méthode non-supervisée d'extraction de termes-clés basée sur TextRank. TopicRank groupe les termes-clés candidats selon leur appartenance à un sujet, représente le document sous la forme d'un graphe complet de sujets, ordonne les sujets selon leur importance, puis sélectionne pour chacun des meilleurs sujets son candidat le plus représentatif. La notion de sujet est vague, tant elle peut exprimer un thème ou un domaine général (par exemple, « le traitement automatique des langues ») ou plus spécifique (par exemple, « l'extraction non-supervisée de termes-clés »). Ici, nous nous intéressons aux sujets les plus spécifiques, car ils caractérisent avec plus de précision le contenu d'un document. Notre approche possède plusieurs avantages, par rapport à TextRank, que nous détaillons ci-dessous :

1) Le regroupement des termes-clés candidats en sujets supprime en amont les problèmes de redondance dans les termes-clés extraits.

2) Le fait d'utiliser des sujets à la place des mots permet de construire un graphe plus compact, de renforcer le poids des arêtes dans le graphe et d'améliorer la qualité de l'ordonnancement.

3) La construction d'un graphe complet permet de supprimer le paramètre de la fenêtre de mots et de capturer de manière plus précise le niveau de relation entre les sujets.

Pour évaluer notre méthode, nous utilisons quatre collections de test aux propriétés différentes (nature des documents, taille des documents, langue, etc.). Nous comparons TopicRank à trois autres méthodes non-supervisées et détaillons l'impact de chacune des contributions que nous proposons.

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 travaux futurs dans la section 5.

2. État de l'art

TopicRank faisant partie des méthodes non-supervisées, l'état de l'art présenté ici se focalise sur cette catégorie de méthodes. L'extraction automatique non-supervisée de termes-clés est une tâche décomposée en général en quatre étapes. Les méthodes non-supervisées traitent les documents généralement un à un. Ils sont tout d'abord enrichis linguistiquement, c'est-à-dire segmentés en phrases, segmentés en mots et étiquetés en parties du discours. Des termes-clés candidats en sont extraits, puis ordonnés afin de ne sélectionner que les plus pertinents (voir la figure 1). L'extraction des termes-clés candidats et leur ordonnancement sont les deux étapes auxquelles nous nous intéressons dans cet article. L'ordonnancement des termes-clés candidats est le

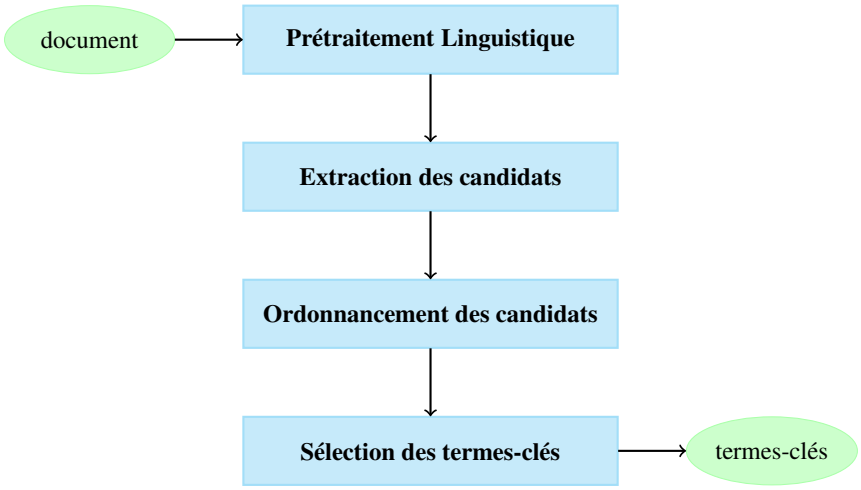


Figure 1. *Les quatre principales étapes de l'extraction automatique de termes-clés.*

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, c'est-à-dire toutes les unités textuelles pouvant être extraites du document analysé, aux seules unités textuelles ayant des particularités similaires à celles des termes-clés tels qu'ils peuvent être donnés par des humains. Il y a deux avantages à cela. Le premier, très évident, est la réduction du temps de calcul nécessaire à l'extraction des termes-clés. Le second avantage est la suppression en amont d'unités textuelles non pertinentes, ces dernières pouvant très négativement affecter les performances de l'ordonnancement. Pour distinguer les différents candidats extraits, nous définissons deux catégories : les candidats positifs, qui sont présents en tant que termes-clés de référence dans nos collections de données, et les candidats non positifs. Parmi les candidats non positifs, nous distinguons deux sous-catégories : les candidats porteurs d'indices de différentes natures pouvant influencer la promotion de candidats positifs (par exemple, la présence des candidats « alerte rouge », « alerte jaune » et « alerte orange » influence l'extraction du candidat positif « alerte » en tant que terme-clé, dans l'article *44960* de notre collection WikiNews – voir la section 4) et les candidats non pertinents, que nous considérons comme des erreurs.

Dans les travaux précédents, trois méthodes d'extraction de candidats sont classiquement utilisées : l'extraction de n-grammes, de chunks nominaux, et d'unités textuelles respectant certains patrons grammaticaux.

Les n-grammes sont toutes les séquences ordonnées de n mots, avec $n \in 1..m$, où m vaut généralement 3 (Witten *et al.*, 1999 ; Turney, 2000 ; Hulth, 2003). Leur extraction est très exhaustive, elle fournit un grand nombre de termes-clés candidats, maximisant ainsi la quantité de candidats positifs, la quantité de candidats porteurs d'indices utiles, mais aussi la quantité de candidats non pertinents. Pour pallier en partie ce problème, il est courant d'utiliser une liste de mots outils pour filtrer les candidats. Les mots outils regroupent les mots fonctionnels de la langue (conjonctions, prépositions, etc.) et les mots courants (« particulier », « près », « beaucoup », etc.). Ainsi, un n-gramme contenant un mot outil en début ou en fin n'est pas considéré comme un terme-clé candidat. Malgré son aspect bruité, ce type d'extraction est encore largement utilisé parmi les méthodes supervisées (Witten *et al.*, 1999 ; Turney, 2000 ; Hulth, 2003). En effet, la phase d'apprentissage de celles-ci les rend moins sensibles aux éventuels candidats erronés (bruit) par rapport aux méthodes supervisées.

Les chunks nominaux sont des syntagmes non récursifs dont la tête est un nom, accompagné de ses éventuels déterminants et modifieurs usuels. Ce sont des segments linguistiquement définis rendant leur extraction plus fiable que celle des n-grammes. 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) constate qu'en tirant parti 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'usage de ce trait supplémentaire a pour effet de filtrer les n-grammes grammaticalement incorrects, favorisant alors l'extraction des candidats positifs, qui sont plus nombreux que ceux parmi les chunks nominaux.

L'extraction d'unités textuelles à partir de patrons grammaticaux prédéfinis permet de contrôler avec précision la nature et la grammaticalité des candidats extraits. À l'instar des chunks nominaux, leur extraction est plus fondée linguistiquement que celle des n-grammes filtrés, et comparée à eux, elle fournit un plus grand nombre de candidats positifs. Dans ses travaux, Hulth (2003) choisie d'extraire des candidats avec les patrons des termes-clés de références les plus fréquents (plus de 10 occurrences) dans sa collection d'apprentissage, tandis que d'autres chercheurs tels que Wan et Xiao (2008) et Hasan et Ng (2010) se concentrent uniquement sur les plus longues séquences de noms (noms propres inclus) et d'adjectifs. Pour des méthodes non-supervisées telles que la nôtre, l'extraction des séquences de noms et d'adjectifs est intéressante, car elle ne nécessite ni données supplémentaires, ni adaptation particulière pour une langue donnée (tel que c'est le cas pour l'extraction des chunks nominaux, par exemple).

Dans le but d'améliorer la qualité des candidats extraits à partir d'articles scientifiques, Kim *et al.* (2009) proposent un filtrage des candidats en fonction de leur spécificité vis-à-vis du document analysé. Cette spécificité est déterminée par rapport à la fréquence d'un candidat dans le document et le nombre de documents d'une collection dans lesquels il apparaît (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. Cette approche est intéressante, mais elle requiert des documents supplémentaires et la définition d'un seuil pour le filtrage. Dans le cas de TopicRank, nous tentons de nous abstraire de l'usage d'autres documents que celui qui est analysé, cette méthode d'extraction de candidats n'est donc pas compatible avec nos objectifs.

2.2. *Ordonnancement des termes-clés candidats*

L'étape d'ordonnancement intervient après l'extraction des termes-clés candidats. Son rôle est de déterminer quels sont, parmi les candidats, les termes-clés d'un document. Les méthodes non-supervisées d'extraction automatique de termes-clés emploient des techniques très différentes, allant du simple usage de mesures fréquentielles (Spärck Jones, 1972 ; Paukkeri et Honkela, 2010) à l'utilisation de modèles de langues obtenus à partir de données non-annotées (Tomokiyo et Hurst, 2003), en passant par 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 décrivons ici uniquement les travaux effectués au sujet de cette catégorie de méthodes.

Mihalcea et Tarau (2004) proposent TextRank, une méthode d'ordonnancement d'unités textuelles à partir d'un graphe. Utilisés dans de nombreuses applications du TAL (Kozareva *et al.*, 2013), les graphes ont l'avantage de présenter de manière simple et efficace les unités textuelles d'un document et les relations qu'elles entretiennent entre elles. De plus, ils bénéficient de nombreuses études théoriques donnant lieu à des outils et algorithmes capables de résoudre de nombreux problèmes du TAL, tels que le résumé automatique (Wan *et al.*, 2007), la compression multi-phrase (Boudin et Morin, 2013) et la désambiguïsassion de texte (Schwab *et al.*, 2013). Dans le cas de TextRank, les nœuds du graphe sont les mots du document et les arrêtes sont leurs relations de co-occurrences dans une fenêtre de 2 mots. Un score d'importance est calculé pour chaque mot à partir de l'algorithme PageRank (Brin et Page, 1998) qui est issu de la mesure de centralité des vecteurs propres. Le principe utilisé est celui de la recommandation, ou du vote, c'est-à-dire un mot est d'autant plus important qu'il co-occurre avec un grand nombre de mots et si les mots avec lesquels il co-occurre sont eux aussi important. Les mots les plus importants sont considérés comme des mots-clés. Ces mots-clés sont marqués dans le document et les plus longues séquences de mots-clés adjacents sont extraites en tant que 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-occurrence. Dans nos travaux, nous créons un graphe dont la qualité ne dépend pas d'un paramètre tel que cette fenêtre de co-occurrence.

Wan et Xiao (2008) modifient TextRank et proposent SingleRank. Dans un premier temps, leur méthode augmente la précision de l'ordonnancement grâce à une pondération des liens de co-occurrence avec le nombre de co-occurrences des mots liés (deux mots qui co-occurrent deux fois, par exemple, sont liés par une arête dont le

poids vaut 2). Dans un second temps, les termes-clés ne sont plus générés, mais ordonnés à partir de la somme du score d'importance des mots qui les composent. Cette nouvelle méthode donne dans la majorité des cas des résultats meilleurs que ceux de TextRank. Cependant, 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 redondants. 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 qui contiennent entre autres « bio-politique » occupent les classements 2 à 8. Dans nos travaux, nous n'ordonnons pas les termes-clés candidats en fonction des mots qu'ils contiennent et évitons ainsi le problème rencontré avec SingleRank.

Toujours dans l'optique d'améliorer l'efficacité de l'ordonnancement, Wan et Xiao (2008) étendent SingleRank en utilisant des documents similaires au document analysé. Sélectionnés à partir d'une vaste collection couvrant plusieurs thématiques, les documents similaires fournissent plus de données relatives aux mots du document analysé et aux relations qu'ils entretiennent. La méthode consiste à utiliser les co-occurrences observées dans les documents similaires pour ajouter ou renforcer des liens dans le graphe. 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 disponibilité de documents similaires à celui qui est analysé.

À l'instar de Wan et Xiao (2008), Tsatsaronis *et al.* (2010) tentent d'améliorer TextRank. Dans leur méthode, ils créent et pondèrent une arrête entre deux mots si et seulement si ceux-ci sont sémantiquement liés selon deux mesures définies à partir de WordNet (Miller, 1995) et de 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 en ajoutant le poids TF-IDF des mots dans le calcul de l'importance des mots, leur méthode est capable de donner de meilleurs résultats que TextRank. Ceci suggère qu'il existe des traits, tels que la première apparition dans le texte et la spécificité des candidats, qui peuvent influencer positivement l'ordonnancement. Dans notre approche, nous estimons que l'ordonnancement des termes-clés candidats vis-à-vis des sujets auxquels ils appartiennent est un moyen de prendre en compte leur spécificité dans le document analysé. En effet, un candidat appartenant à un sujet important d'un document est intuitivement plus spécifique à ce document qu'un candidat appartenant à un sujet moins important.

L'usage de sujets dans le processus d'ordonnancement de 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 ces différents ordonnancements afin d'obtenir un rang global pour chaque mot. 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 pas le modèle LDA.

De plus, il nous semble plus judicieux, d'un point de vue complexité, d'effectuer un seul ordonnancement.

3. Extraction de termes-clés avec TopicRank

TopicRank est une méthode non-supervisée d'extraction de termes-clés qui modélise un document 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. Dans un premier temps, les sujets sont identifiés, puis dans un second temps, ordonnés. Enfin, les candidats les plus représentatifs des sujets les plus importants sont extraits comme termes-clés.

3.1. *Identification des sujets*

La première étape de l'identification des sujets consiste à extraire les termes-clés candidats. Afin de réaliser une identification de qualité des sujets, nous excluons l'extraction des n-grammes qui fournie beaucoup plus de candidats non pertinents que les autres méthodes. Parmi l'extraction des chunks nominaux et l'extraction des plus longues séquences de noms et d'adjectifs, nous suivons Wan et Xiao (2008) et Hasan et Ng (2010) en choisissant d'extraire les plus longues séquences de noms et d'adjectifs. Cette méthode présente l'avantage de fournir plus de candidats positifs que l'extraction des chunks nominaux tout en n'introduisant pas un nombre significatif de candidats non pertinents. De plus, l'extraction des plus longues séquences de noms et d'adjectifs fournit des candidats grammaticalement corrects et requiert une adaptation limitée pour le traitement de documents d'une autre langue, telle que le français ou l'anglais.

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. Ceux-ci sont groupés en fonction d'une similarité de Jaccard (voir l'équation 1) dans laquelle ils sont considérés comme des sacs de mots, les mots étant tronqués selon la méthode de Porter (1980) afin de considérer identiques les variantes flexionnelles et dérivationnelles. 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. À cela s'ajoute aussi des erreurs introduites par la méthode de Porter (1980) (par exemple, « empire » et « empirique » partagent le même radical, « empir »).

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

Une fois la similarité connue entre toutes les paires de candidats, nous appliquons l'algorithme de groupement hiérarchique agglomératif (*Hierarchical Agglomerative Clustering – HAC*). Initialement, chaque candidat représente un groupe et, 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 déterminée à partir de la similarité de Jaccard calculée entre les candidats de chaque groupe. Il existe trois stratégies pour calculer la similarité entre deux groupes :

– 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).

L'une ou l'autre de ces stratégies est à privilégier en fonction du type des candidats qui sont extraits. Pour des candidats qui ont de forts recouvrements, tels que les n-grammes, il semble par exemple plus pertinent d'utiliser la stratégie complète qui est la moins agglomérative. Dans le cas de TopicRank, les termes-clés candidats étant de meilleure qualité que les n-grammes, la stratégie moyenne est une meilleure alternative.

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 cette pondération, nous choisissons d'utiliser la force du lien 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, mais nous utilisons la distance entre les candidats des sujets dans le document :

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

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

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.

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

Une fois le graphe construit, l'algorithme d'ordonnancement de TextRank est utilisé pour identifier quels sont les sujets les plus importants du document. Cet ordonnancement se fonde sur le principe de recommandation, ou de vote, c'est-à-dire un sujet est d'autant plus important qu'il est fortement connecté avec un grand nombre de sujets et que les sujets avec lesquels il est fortement connecté sont importants :

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

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

3.3. *Sélection des termes-clés*

La sélection des termes-clés est la dernière étape de TopicRank. Elle consiste à chercher les termes-clés candidats qui représentent le mieux les sujets importants qui sont abordés dans le document. Dans le but de ne pas extraire de termes-clés redondants, un seul candidat est sélectionné par sujet. Ainsi, pour k sujets, k termes-clés non redondant couvrant exactement k sujets sont extraits.

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 a la plus haute similarité avec les autres candidats du sujet (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 au sujet. 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

2. $V_i = \{v_i \mid \forall v_j in V, v_j \neq v_i\}$, car G est un graphe complet.

groupe risque d'avoir un effet similaire, car « réseau » est le sous-composant de nombreux autres candidats tels que « réseau étendu » (*wide area network*), « réseaux locaux » (*local area networks*), « réseaux informatisés de communication » (*computer-communication networks*).

La figure 2 donne un exemple d'extraction de termes-clés avec TopicRank. Dans cet exemple, 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 apparaissent dans la référence.

4. Évaluation

Pour valider notre approche, nous réalisons une première série d'évaluations visant à déterminer la configuration optimale de TopicRank. Nous comparons ensuite TopicRank aux méthodes précédentes et analysons l'impact de chacune de nos contributions.

4.1. *Cadre expérimental*

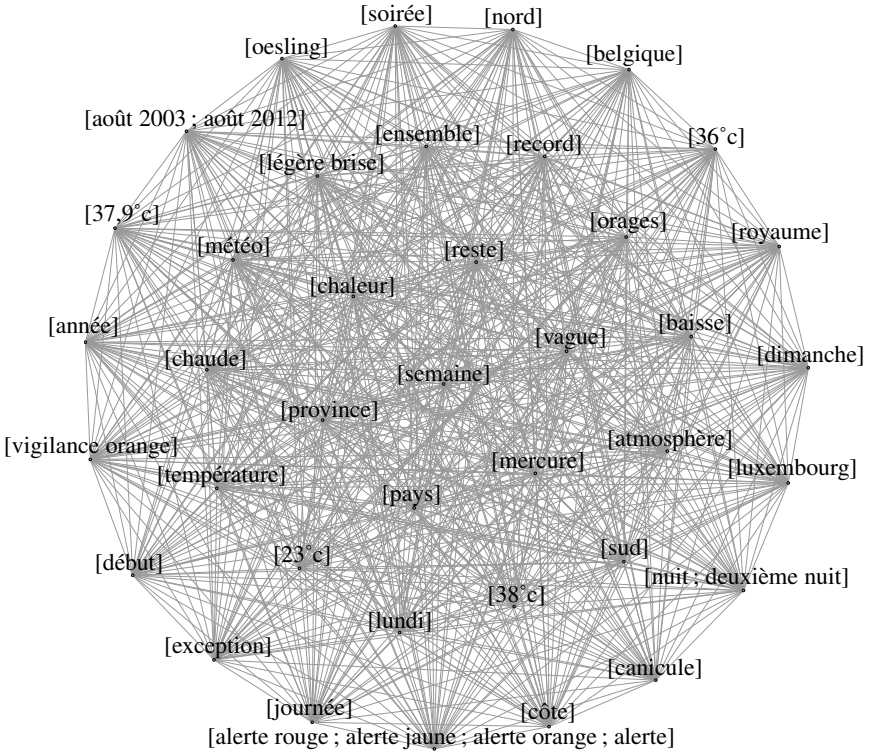
4.1.1. *Données de test*

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 ici diffèrent 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 ont été annotés par Wan et Xiao (2008). Lors de nos expériences, nous utilisons ces 308 documents.

SemEval (Kim *et al.*, 2010) est la collection en anglais 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. 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

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.*

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.*

documents, ils correspondent à la combinaison de ceux donnés par les auteurs et des lecteurs.

WikiNews[3] est une collection de 100 articles journalistiques en français que nous avons extraits 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 en 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 des 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 langue une collection 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 associés varie, de même que le nombre de termes-clés n'apparaissant pas dans les documents.

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

4.1.2. *Prétraitement*

Chaque document des collections de données utilisées subit les mêmes prétraitements. Chaque document est tout d'abord segmenté en phrases, puis en mots et enfin étiqueté 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. Tous ces outils sont utilisés avec leur configuration par défaut.

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 ne pas considérer fausse l'extraction d'une variante flexionnelle d'un terme-clé de référence, les opérations de comparaison sont effectuées à partir de la forme radicale des mots.

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 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 de référence sont implémentées par nous-même. Lorsque celles-ci ont des parties communes avec TopicRank, elles bénéficient des mêmes composants. De plus, pour améliorer les résultats des méthodes de référence, leurs sorties sont filtrées afin de supprimer les termes-clés candidats dont les mots ont la même forme radicale que ceux d'un candidat mieux classé. Ce filtrage ne dégrade en rien les résultats et a pour effet de faire monter de nouveaux candidats dans le classement.

4.2. *Analyse 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 la configuration de groupement optimale (variation du seuil ζ et de la stratégie de groupement) et l'autre pour déterminer la stratégie de sélection des termes-clés représentatifs.

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

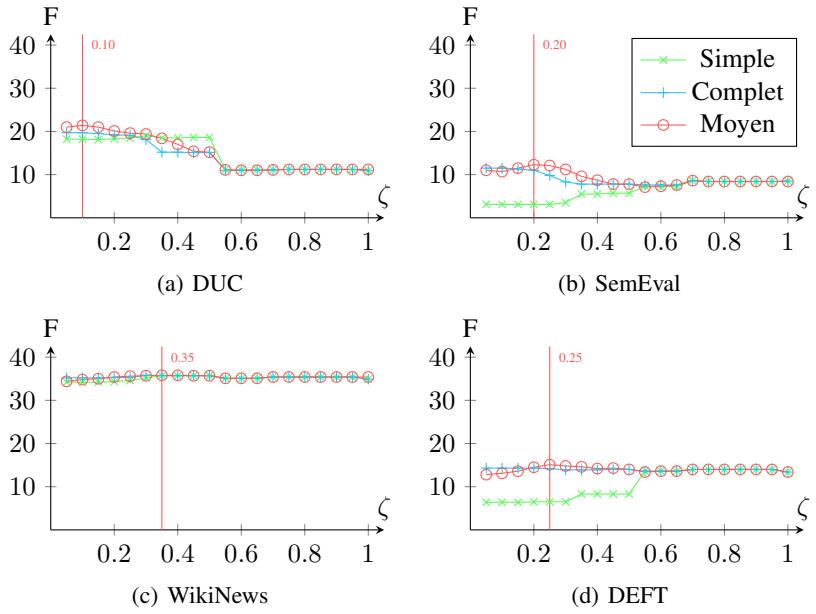


Figure 3. *Résultats de l'extraction de 10 termes-clés, avec TopicRank, en fonction de la stratégie de regroupement et de la valeur du seuil de similarité* ζ.

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 groupement. La stratégie de sélection d'un terme-clé 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 prenne 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. Nous observons tout de même 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, de variantes dérivationnelles où de candidats dont ceux contenant le plus de mots incluent 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 en plus « églises évangéliques », ainsi que d'autres, lorsque ζ est

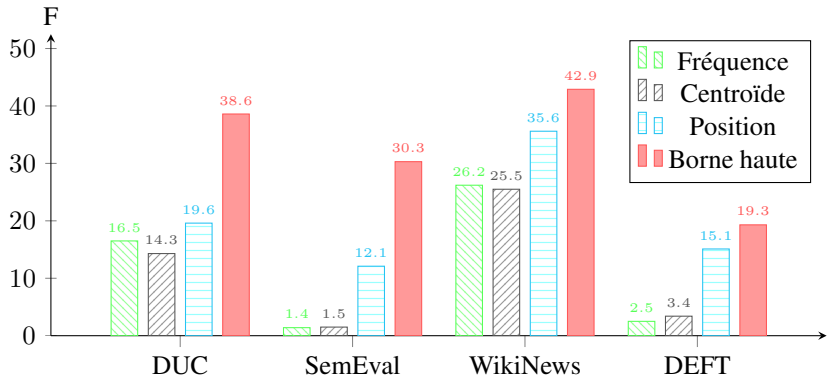


Figure 4. *Résultats de l'extraction de 10 termes-clés, avec TopicRank, en fonction des différentes sélections de termes-clés candidats par sujet.*

plus faible). Après observation des résultats de cette expérience, le seuil ζ est fixé à 0,25 pour toutes les expériences suivantes. De même, la stratégie de groupement utilisée dans la suite est la stratégie moyenne.

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, en effet, 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 satisfaisants, nous remarquons qu'il existe encore une marge de progression importante. Les valeurs indiquées par la borne haute représentent les résultats qui pourraient être obtenus avec un oracle. Pour chacun des sujets les plus importants, l'oracle sélectionne toujours un candidat positif, s'il y en a un. La 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 méthodes de référence. Globalement, TopicRank donne de meilleurs résultats que les méthodes de référence utilisées. Comparée à la méthode TF-IDF, TopicRank donne de meilleurs résultats pour SemEval, WikiNews et DEFT. Cette supériorité vis-à-vis de TF-IDF est importante à noter, car cette méthode obtient de bons résultats en tirant parti de statistiques extraites de documents supplémentaires (apprentissage non-supervisé), alors que TopicRank n'utilise que le document à analyser. Comparée aux autres méthodes à base de graphe, TopicRank donne des résultats significativement

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.*

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.*

meilleurs pour SemEval, WikiNews et DEFT. Ceci confirme donc que le groupement des candidats permet de rassembler des informations améliorant la précision de l'ordonnancement. En ce qui concerne DUC, notre méthode est toujours significativement meilleure que TextRank, mais elle ne l'est pas vis-à-vis des autres méthodes. Aux vues de la borne haute de la figure 4, l'une des raisons pour lesquelles les résultats sont moins bons pour DUC est que la stratégie de sélection des candidats les plus représentatifs des sujets n'est pas adaptée. En effet, la différence avec la borne haute est de 19,0 points de f-score. Une analyse plus approfondie des différents apports de TopicRank peut aussi donner une piste sur les raisons de ces moins bons résultats.

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 à la place des mots (respectivement +complet, +candidats et +sujets). Les résultats de ces trois variantes de SingleRank sont présentés 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.

Cela confirme la pertinence d'ordonner directement les candidats, plutôt que les mots. De plus, le groupement des candidats représentant le même sujet améliore la précision de l'ordonnancement grâce à la mutualisation des relations qu'ils entretiennent avec les candidats représentant d'autres sujets. L'usage d'un graphe complet, quant à lui, n'améliore pas significativement les résultats de SingleRank. Ceux-ci sont compétitifs vis-à-vis de ceux obtenus en construisant un graphe de co-occurrences. Nous pensons tout de même que l'usage du graphe complet est à privilégier afin d'éviter la fenêtre de co-occurrences.

En ce qui concerne la collection DUC, le tableau 3 montre une perte de performance induite par la construction du graphe avec les termes-clés candidats. Cette perte de performance s'explique par le fait qu'il y a, dans les documents de DUC, peu de répétition des candidats, notamment ceux de plus d'un mot. Le graphe créé contient alors moins de relations de co-occurrences que lorsque les nœuds sont les mots du document et est donc moins précis pour l'ordonnancement. Ceci se confirme avec la figure 4 qui montre que l'extraction des candidats les plus fréquents, pour chaque sujet de DUC, donne des résultats proches de ceux obtenus avec l'extraction des candidats apparaissant en premier dans le document.

4.4. *Analyse des sujets détectés*

Dans cette section, nous analysons les groupements en sujets effectués par TopicRank et tentons de déterminer quelles sont les causes principales d'erreurs. Notre langue maternelle étant le français et les documents de WikiNews étant de trop courts pour que nous puissions observer un nombre significatif de groupements en sujets, nos observations ci-dessous sont faites à partir des documents de la collection DEFT. Nous distinguons deux types d'erreurs : celles qui sont liées à du bruit introduit dans les groupes et celles qui sont liées à des candidats ayant des propriétés particulières.

Nous observons des erreurs liées à l'extraction des termes-clés candidats. Lors de l'extraction des candidats, certaines unités textuelles sont extraites à cause d'erreurs dans l'étiquetage grammatical. Ces erreurs concernent principalement la détection des prépositions composées et la détection des participes. Dans le document *as_2006_014935ar*, nous observons par exemple que « pays dits » est un terme-clé candidats, car le participe passé « dits » est considéré comme un adjectif dans la phrase « [...] elles ne cessent de se développer à travers le monde et particulièrement dans les pays dits “du sud” [...] ».

En ce qui concerne les candidats ayant des propriétés particulières, nous observons tout d'abord de nombreuses erreurs lorsque les groupements sont déclenchés par un adjectif. Ce sont particulièrement les expansions nominales s'effectuant à gauche qui sont la source d'erreurs (« même langue » groupé avec « même représentation », « grands traits » groupé avec « grande ignorance », etc.). Parmi les expansions nominales s'effectuant à droite, les adjectifs relationnels sont moins sujets aux erreurs que les autres adjectifs. Notons tout de même que lorsque ces adjectifs ont attraits au

contexte général du document, ils sont très fréquemment utilisés et beaucoup de candidats les contenant sont groupés par erreur. Ainsi, dans le document *as_2002_000707ar* qui examine l'organisation du dialogue politique économique entre les membres de la commission européenne, les termes-clés candidats « forces économiques », « type économique », « délabrement économique » et « économies postsocialistes » sont groupés par erreur, car ils partagent tous l'adjectif « économique ». Outre ces groupements erronés, nous observons aussi de mauvais groupements lorsque les candidats ne contiennent que très peu de mots. En effet, pour les candidats de deux mots, il ne suffit que d'un seul mot en commun pour les grouper. Ces candidats étant très fréquents, ils sont la cause de nombreuses erreurs.

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'indices utiles auparavant éparpillés. 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 redondants – pour k termes-clés extraits, exactement k sujets sont couverts. Finalement, le graphe est désormais complet et ne requiert plus le paramètre de fenêtre de co-occurrences.

Les bons résultats de notre méthode montrent la pertinence d'un groupement 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 améliorés (pour un gain maximum allant de 4,2 à 19,0 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. Il a plusieurs limitations liées au calcul de similarité employé pour déterminer si deux termes-clés candidats expriment le même sujet. En effet, ce calcul ne tient compte ni de la synonymie, ni de l'ambiguïté des mots. Aussi, l'usage du radical des mots n'est pas sans introduire du bruit lié à certains faux positifs (par exemple, « empire » et « empirique »), mais aussi à certains faux négatifs (par exemple, « étude » et « étudier »). L'ajout de connaissances concernant les synonymes permettrait de créer des sujets plus complets et la désambiguïsassion éviterait un groupement systématique des termes-clés candidats ayant un ou plusieurs mots en commun. Quant à l'usage de la méthode de

Porter (1980), il est envisagé de le comparer avec l'usage d'une méthode de lemmatisation. D'un point de vue 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 détaillée 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.

Boudin F., Morin E., « Keyphrase Extraction for N-best Reranking in Multi-Sentence Compression », *Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL)*, Association for Computational Linguistics, Atlanta, Georgia, p. 298-305, June, 2013.

Brin S., Page L., « The Anatomy of a Large-Scale Hypertextual Web Search Engine », *Computer Networks and ISDN Systems*, vol. 30, nº 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.

D'Avanzo E., Magnini B., « A Keyphrase-Based Approach to Summarization : the LAKE System at DUC-2005 », *Proceedings of DUC 2005 Document Understanding Conference*, 2005.

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.

Kozareva Z., Matveeva I., Melli G., Nastase V. (eds), *Proceedings of TextGraphs-8 Graph-Based Methods for Natural Language Processing*, Association for Computational Linguistics, Seattle, Washington, USA, October, 2013.

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, n° 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, n° 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, nº 3, p. 130-137, 1980.

Schwab D., Goulian J., Tchechmedjiev A., « Désambiguïsation lexicale de textes : efficacité qualitative et temporelle d'un algorithme à colonies de fourmis », *Traitement Automatique des Langues (TAL)*, 2013.

Spärck Jones K., « A Statistical Interpretation of Term Specificity and its Application in Retrieval », *Journal of Documentation*, vol. 28, nº 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, nº 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.

Wan X., Yang J., Xiao J., « Towards an Iterative Reinforcement Approach for Simultaneous Document Summarization and Keyword Extraction », *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, Association for Computational Linguistics, Prague, Czech Republic, p. 552-559, June, 2007.

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.