HAL open science

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

Adrien Bougouin, Florian Boudin

To cite this version:

Adrien Bougouin, Florian Boudin. TopicRank : ordonnancement de sujets pour l'extraction automatique de termes-clés. Revue TAL : traitement automatique des langues, 2014, pp.45-69. hal-01096913

HAL Id: hal-01096913

https://hal.science/hal-01096913v1

Submitted on 18 Dec 2014

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

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 en sujets, ordonne les sujets et extrait de chacun des meilleurs sujets le terme-clé candidat qui le représente le mieux. 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 non supervisée 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 cannot 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.*

1. Introduction

Un terme-clé, couramment appelé mot-clé, est un mot ou une expression polylexicale permettant de caractériser une partie du contenu d'un document. Groupés ensemble, les termes-clés d'un document permettent de définir les principaux sujets (concepts) abordés dans un document (cf. exemple figure 1). Ils sont alors utiles dans de nombreuses applications du traitement automatique des langues (TAL), telles que le résumé automatique (D'Avanzo et Magnini, 2005), la compression multiphrase (Boudin et Morin, 2013), la classification de documents (Han *et al.*, 2007), ou l'indexation automatique de documents (Medelyan et Witten, 2008), qui nous intéresse plus particulièrement. Pourtant, de nombreux documents, tels que ceux disponibles sur Internet, n'en sont pas accompagnés et la quantité de documents 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, en témoigne la quantité grandissante de travaux scientifiques à ce sujet (Hasan et Ng, 2014) ainsi que l'émergence de campagnes d'évaluation des méthodes d'extraction automatique de termes-clés (Kim *et al.*, 2010 ; Paroubek *et al.*, 2012).

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.

Figure 1. *Exemple de termes-clés mis en évidence dans un article journalistique du site Web WikiNews (`http://fr.wikinews.org/w/index.php?oldid=426443`)*

L'extraction automatique de termes-clés consiste à sélectionner dans un document les unités textuelles les plus importantes parmi un ensemble de termes-clés candidats sélectionnés dans le document. Les termes-clés candidats doivent être de nature similaire à celle des termes-clés assignés par des humains. Ils sont sélectionnés à partir d'hypothèses simples (contiennent des noms, contiennent des adjectifs, etc.) et ne remplissent pas nécessairement les critères d'un terme tel que défini en extraction terminologique. 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 méthodes supervisées tirent profit

d'une collection de documents annotés en termes-clés. Elles apprennent, en amont, un modèle de classification binaire permettant, par la suite, de déterminer quels sont parmi les termes-clés candidats d'un document ceux qui sont des termes-clés et ceux qui n'en sont pas. Les méthodes non supervisées, quant à elles, attribuent un score d'importance à chaque terme-clé candidat en fonction de divers indicateurs, tels que la fréquence, les relations de cooccurrences ou la position dans le document. Du fait de leur phase d'apprentissage, les méthodes supervisées sont en général plus performantes que les méthodes non supervisées. Cependant, la faible quantité de documents annotés en termes-clés disponibles couplée à la forte dépendance des modèles de classification vis-à-vis du type des documents à partir desquels ils sont appris, poussent les chercheurs à s'intéresser de plus en plus aux méthodes non supervisées (Hasan et Ng, 2010).

Les méthodes d'extraction de termes-clés non supervisées les plus étudiées sont sans conteste celles fondées sur TextRank (Mihalcea et Tarau, 2004), qui est une méthode d'ordonnancement d'unités textuelles à partir d'un graphe. Un graphe est un moyen naturel de représenter les unités textuelles et les relations qu'elles entretiennent et de nombreuses applications du TAL en font usage (Kozareva *et al.*, 2013). Dans le cadre de l'extraction de termes-clés, le principe est de représenter le document sous la forme d'un graphe dans lequel les nœuds correspondent aux mots et les arêtes à leurs relations de cooccurrences dans une fenêtre de mots. Un score d'importance est alors calculé pour chaque mot selon le principe de recommandation : un mot est d'autant plus important s'il cooccurre avec un grand nombre de mots et si les mots avec lesquels il cooccurre sont eux aussi importants. Enfin, les mots les plus importants servent à générer des termes-clés pour le document.

Dans cet article, nous présentons TopicRank[1], une méthode non supervisée d'extraction de termes-clés fondée 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 par importance selon le principe de recommandation, puis sélectionne pour chacun des meilleurs sujets le terme-clé candidat qui le représente le mieux. La notion de sujet est vague tant elle peut exprimer un thème, ou un domaine, général (par exemple « traitement automatique des langues ») ou plus spécifique (par exemple « extraction non supervisée de termes-clés »). Ici, nous considérons comme sujet toute information véhiculée par au moins une unité textuelle du document analysé. Notre approche possède plusieurs avantages, en comparaison avec TextRank, que nous détaillons ci-dessous :

– regroupement des termes-clés candidats en sujets supprime en amont les problèmes de redondance dans les termes-clés extraits ;

– usage de sujets à la place de 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 ;

– construction d'un graphe complet permet de supprimer le paramètre de la fenêtre

1. Cette article est une version étendue de (Bougouin *et al.*, 2013).

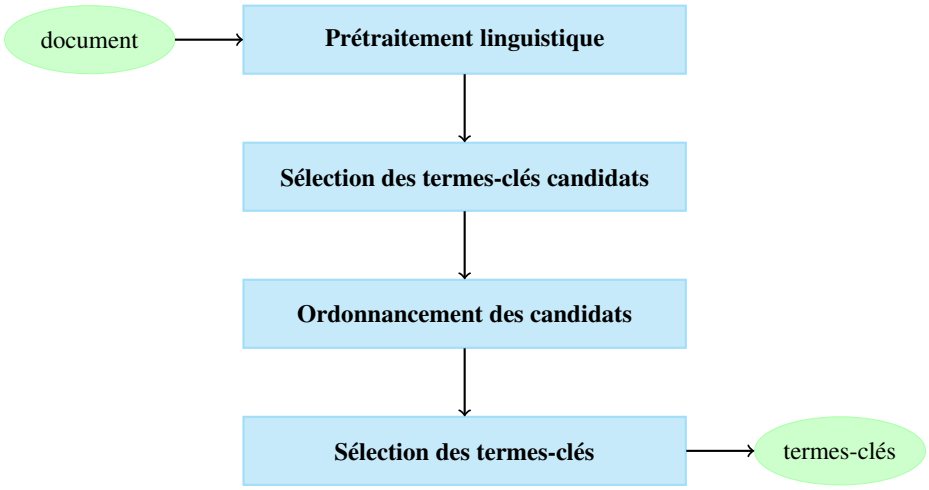


Figure 2. *Les quatre principales étapes de l'extraction automatique de termes-clés*

de cooccurrences et de capturer de manière plus précise le niveau de relations 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 non supervisées 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 analysons les erreurs de TopicRank dans la section 5, puis nous concluons et discutons des travaux futurs dans la section 6.

2. État de l'art

Dans cet état de l'art, nous nous focalisons sur la catégorie de méthodes à laquelle appartient TopicRank, c'est-à-dire les méthodes non supervisées. L'extraction automatique non supervisée de termes-clés est une tâche répartie généralement en quatre étapes. Les méthodes non supervisées traitent les documents un à un. Ceux-ci sont tout d'abord enrichis linguistiquement, c'est-à-dire segmentés en phrases, segmentés en mots et étiquetés grammaticalement. Des termes-clés candidats y sont sélectionnés, puis ordonnés afin de n'extraire que les plus pertinents (voir la figure 2). La sélection des termes-clés candidats et leur ordonnancement sont les deux étapes auxquelles nous nous intéressons dans cet état de l'art. Si l'ordonnancement des termes-clés candidats est le cœur de la tâche d'extraction de termes-clés, ses performances dépendent de

la qualité des candidats préalablement sélectionnés (Wang *et al.*, 2014) et il est donc important de bien choisir leur méthode de sélection.

2.1. *Sélection des termes-clés candidats*

L'objectif de la sélection des termes-clés candidats est de déterminer quelles sont les unités textuelles qui sont potentiellement des termes-clés, c'est-à-dire les unités textuelles qui ont des particularités similaires à celles des termes-clés définis par des humains. Nous savons par exemple que les termes-clés sont majoritairement constitués de noms et d'adjectifs. Cette étape de sélection des termes-clés présente deux avantages. Le premier est la réduction du temps de calcul nécessaire à l'extraction des termes-clés. Le second est la suppression d'unités textuelles non pertinentes pouvant affecter négativement les performances de l'ordonnancement. Pour distinguer les différents candidats sélectionnés, nous définissons deux catégories : les candidats positifs, qui correspondent aux termes-clés assignés par des humains (termes-clés de référence), 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'exemple de la figure 1) et les candidats non pertinents, que nous considérons comme des erreurs.

Dans les travaux précédents, trois méthodes d'extraction de candidats sont utilisées : l'extraction de n-grammes, de *chunks* nominaux, et d'unités textuelles respectant certains patrons grammaticaux. Dans cette section, nous ne présentons que les travaux utilisés pour la sélection des termes-clés candidats. Des travaux connexes pourraient toutefois être considérés : extraction terminologique (Castellví *et al.*, 2001), détection de collocation (Pearce, 2002), etc.

Les n-grammes sont toutes les séquences ordonnées de n mots adjacents. Leur extraction est très exhaustive, elle fournit un grand nombre de termes-clés candidats, maximisant 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 un anti-dictionnaire pour filtrer les candidats. Ce dernier regroupe les mots fonctionnels de la langue (conjonctions, prépositions, etc.) et les mots courants (« particulier », « près », « beaucoup », etc.). Un n-gramme contenant un mot présent dans l'anti-dictionnaire 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). La phase d'apprentissage de celles-ci les rend moins sensibles aux éventuels candidats erronés (bruit) que les méthodes non supervisées.

Exemples de $\{1..3\}$-grammes sélectionnés dans la phrase *« Une légère brise de côte pourra faiblement rafraichir l'atmosphère »* dans l'exemple de la figure 1 : « légère », « brise », « côte », « pourra », « faiblement », « rafraichir », « atmosphère », « lé-

gère brise », « côte pourra », « pourra faiblement », « faiblement rafraichir », « brise de côte », « côte pourra faiblement », « pourra faiblement rafraichir » et « rafraichir l'atmosphère ».

Les *chunks* nominaux sont des syntagmes non récursifs dont la tête est un nom accompagné de ses éventuels déterminants et modifieurs usuels. Ils sont linguistiquement définis et donc plus fiables que les n-grammes, comme le montrent les expériences menées par Hulth (2003) et Eichler et Neumann (2010). Cependant, Hulth (2003) constate que l'usage de l'étiquetage grammatical des termes-clés candidats lors de l'extraction supervisée de termes-clés permet d'éliminer les n-grammes grammaticalement incorrects et d'obtenir de meilleurs performances qu'avec les *chunks* nominaux. Contrairement à Hulth (2003), nous proposons une méthode d'extraction non supervisée de termes-clés. Sélectionner les *chunks* nominaux est donc une solution plus fiable que de sélectionner les n-grammes.

Exemples de *chunks* nominaux sélectionnés dans la phrase *« Une légère brise de côte pourra faiblement rafraichir l'atmosphère »* dans l'exemple de la figure 1 : « une légère brise », « côte » et « l'atmosphère ».

La sélection d'unités textuelles à partir de patrons grammaticaux prédéfinis permet de contrôler avec précision la nature et la grammaticalité des candidats à sélectionner. À l'instar des *chunks* nominaux, leur sélection est plus fondée linguistiquement que celle des n-grammes. Dans ses travaux, Hulth (2003) sélectionne les candidats à partir des patrons des termes-clés de référence les plus fréquents dans sa collection d'apprentissage (plus de dix occurrences), 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, la sélection des séquences de noms et d'adjectifs est intéressante, car elle ne nécessite ni données supplémentaires, ni adaptation particulière pour une langue donnée, c'est le cas pour les *chunks* nominaux.

Exemples de plus longues séquences de noms et d'adjectifs sélectionnées dans la phrase *« Une légère brise de côte pourra faiblement rafraichir l'atmosphère »* dans l'exemple de la figure 1 : « légère brise », « côte » et « atmosphère ».

2.2. *Ordonnancement des termes-clés candidats*

L'étape d'ordonnancement intervient après la sélection des termes-clés candidats. Son rôle est de déterminer quels sont parmi les candidats d'un document ceux qui sont les plus importants. Les méthodes non supervisées d'extraction automatique de termes-clés emploient des techniques très différentes, allant du simple usage de mesures fréquentielles (Paukkeri et Honkela, 2010) à l'utilisation de modèles de langues (Tomokiyo et Hurst, 2003), en passant par la construction d'un graphe de cooccurrences (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 pour cette catégorie de méthode.

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. Dans le cas de TextRank, les nœuds du graphe sont les mots du document et leurs arêtes représentent leurs relations d'adjacence dans le document, c'est-à-dire leurs relations de cooccurrences 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 (du vote) : un mot est d'autant plus important s'il cooccurre avec un grand nombre de mots et si les mots avec lesquels il cooccurre sont eux aussi importants. 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. Bien qu'elle utilise une représentation intéressante et efficace d'un document, cette méthode présente l'inconvénient d'ordonner uniquement les mots plutôt que les termes-clés candidats. Dans nos travaux, nous proposons d'ordonner directement les termes-clés candidats.

Wan et Xiao (2008) modifient TextRank et proposent SingleRank. Dans un premier temps, leur méthode augmente la précision de l'ordonnancement en utilisant une fenêtre de cooccurrences élargie (empiriquement) à dix et en pondérant les arêtes par le nombre de cooccurrences entre les deux mots qu'elles relient. Dans un second temps, les termes-clés ne sont plus générés, mais ordonnés à partir de la somme des scores d'importance des mots qui les composent. Cette nouvelle méthode donne, dans la majorité des cas, des résultats meilleurs que ceux de TextRank. Cependant, la précision de l'ordonnancement dépend de la valeur de la fenêtre de cooccurrences qui est fixée, en témoignent les observations contradictoires de Mihalcea et Tarau (2004) et de Wan et Xiao (2008) lorsqu'ils appliquent leurs méthodes avec différentes fenêtres sur des documents de natures différentes. De plus, faire la somme des scores d'importance des mots pour ordonner les candidats est une approche qui a pour conséquence de favoriser les plus longues séquences tout en faisant monter dans le classement des candidats redondants (par exemple dans l'exemple de la figure 1, le candidat positif « alerte » est classé quatrième par SingleRank, alors que les candidats non positifs « alerte orange », « alerte jaune » et « alerte rouge » qui le contiennent occupent de meilleurs classements). Ici, nous proposons d'utiliser un graphe complet pour éviter de définir une fenêtre de cooccurrences et de grouper les termes-clés candidats afin d'éviter le problème de redondance.

Toujours dans le but d'améliorer l'efficacité de l'ordonnancement proposé par Mihalcea et Tarau (2004), Wan et Xiao (2008) étendent SingleRank en utilisant des documents similaires au document analysé, selon la mesure de similarité vectorielle cosinus. Faisant l'hypothèse que ces documents similaires fournissent des données supplémentaires relatives aux mots du document analysé et aux relations qu'ils entretiennent, ils utilisent les relations de cooccurrences 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. Toutefois, ses performances sont fortement

liées à la disponibilité de documents similaires à celui qui est analysé. Cette méthode ne peut donc être appliquée que dans un contexte particulier, contexte que nous ne pouvons garantir dans ce travail.

Tsatsaronis *et al.* (2010) tentent eux aussi d'améliorer TextRank. Dans leur méthode, ils créent et pondèrent une arête entre deux mots si et seulement si ceux-ci sont sémantiquement liés dans WordNet (Miller, 1995) ou dans Wikipedia (Milne et Witten, 2008). Leurs expériences 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 (Spärck Jones, 1972) des mots 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 de TextRank est à l'origine proposé par Liu *et al.* (2010). Reposant sur un modèle LDA (*Latent Dirichlet Allocation*) (Blei *et al.*, 2003), 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, dans le but de proposer une méthode générique qui ne requiert aucun travail préalable, nous tentons de nous abstraire de l'usage de documents supplémentaires et n'utilisons pas le modèle LDA.

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. Notre méthode repose sur les trois étapes suivantes qui seront détaillées dans la suite : identification des sujets ; ordonnancement des sujets ; sélection des termes-clés.

3.1. *Identification des sujets*

Dans ce travail, un sujet est une information générale ou le plus souvent spécifique véhiculée par au moins une unité textuelle présente dans le document analysé. Nous ne nous reposons donc pas sur une formulation approfondie d'un sujet, nous groupons seulement les termes-clés candidats lorsque ceux-ci véhiculent la même information.

La première étape de l'identification des sujets consiste à sélectionner les termes-clés candidats. Afin de réaliser une identification de qualité des sujets, nous excluons la sélection des n-grammes qui fournit beaucoup plus de candidats non pertinents que les autres méthodes. Concernant le choix entre la sélection des *chunks* nominaux et la sélection des plus longues séquences de noms et d'adjectifs, nous suivons Wan et

Xiao (2008) et Hasan et Ng (2010) en sélectionnant les plus longues séquences de noms et d'adjectifs. Cette méthode présente l'avantage de fournir des candidats grammaticalement corrects et de ne nécessiter qu'une adaptation limitée pour le traitement de documents d'une autre langue.

La seconde étape de l'identification des sujets consiste à grouper les termes-clés candidats lorsqu'ils appartiennent au même sujet. Dans le souci de proposer une méthode ne faisant pas l'usage de données supplémentaires, nous optons pour un groupement quelque peu naïf des candidats. Deux candidats c_1 et c_2 sont groupés en fonction d'une similarité de Jaccard par laquelle ils sont considérés comme des sacs de mots tronqués selon la méthode de racinisation[2] de Porter (1980) :

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

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 l'usage de la méthode de Porter (1980) (par exemple les mots « empire » et « empirique » partagent le même radical, « empir »).

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, jusqu'à l'obtention d'un nombre prédéfini de groupes, 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 extraits. Pour des candidats qui ont de forts recouvrements, tels que les n-grammes, il semble plus pertinent d'utiliser la stratégie complète qui est la moins agglomérative. Dans le cas de TopicRank, où les candidats sont de meilleure qualité que les n-grammes, la stratégie moyenne est une meilleure alternative.

2. Cette racinisation a pour effet de grouper les candidats qui varient uniquement en termes de flexion ou de dérivation.

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 $G = (V, E)$, où E est l'ensemble des liens entre les nœuds[3]. Du fait que le graphe utilisé soit 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 suivons Wan et Xiao (2008) et utilisons la force du lien sémantique entre les sujets. Cependant, parce que nous utilisons un graphe complet, il ne nous est pas possible de représenter cette force par le nombre de cooccurrences entre les sujets. Pour préserver l'intuition derrière l'usage du nombre de cooccurrences, nous représentons la force du lien sémantique entre deux sujets par la distance entre les candidats des sujets dans le document :

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

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

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 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 (de vote), c'est-à-dire un sujet est d'autant plus important s'il est fortement connecté avec un grand nombre de sujets et si les sujets avec lesquels il est fortement connecté sont importants :

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

où V_i est l'ensemble des sujets connectés au sujet[4] s_i et où λ est un facteur d'atténuation définit à 0,85 d'après les recommandations de 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.

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

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 redondants 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 proposons trois stratégies de sélection pouvant répondre à ce problème :

– la première position : en supposant qu'un sujet est tout d'abord introduit sous 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 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. Sélectionner les candidats les plus fréquents risque de ne pas être une solution stable selon les types de documents, en particulier selon leur taille, tandis que sélectionner les centroïdes risque de ne pas fournir les termes-clés les plus précis.

La figure 3 donne un exemple d'extraction de termes-clés avec TopicRank à partir de l'exemple de la figure 1. Dans cet exemple, nous observons un groupement correct de toutes les variantes de « alertes », mais aussi un groupement erroné de « août 2003 » avec « août 2012 ». 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 l'ensemble de termes-clés assignés par des humains pour ce document.

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 travaux précédents et analysons l'impact de chacune de nos contributions.

4.1. *Cadre expérimental*

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

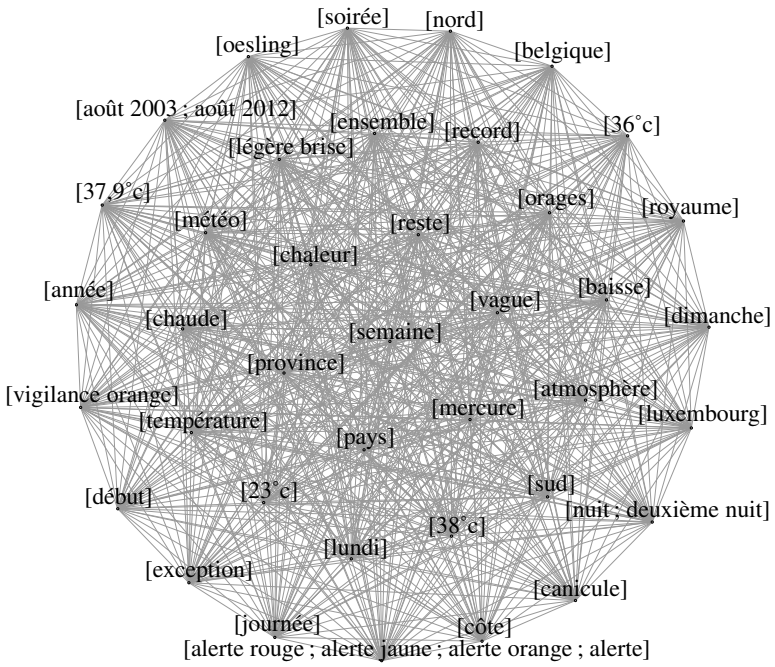
Dans nos expérimentations, nous comparons TopicRank à trois autres méthodes non supervisées d'extraction automatique de termes-clés. Nous choisissons TextRank et SingleRank, les deux méthodes qui sont la fondation des méthodes à base de graphe,

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 3. *Exemple d'extraction de termes-clés avec TopicRank. Les termes-clés soulignés sont les termes-clés correctement extraits.*

et la méthode de pondération TF-IDF. Cette dernière méthode consiste à extraire en tant que termes-clés les candidats dont les mots sont importants. Un mot est considéré important dans un document s'il y est fréquent (TF élevé) et s'il a une forte spécificité (IDF élevé) calculée à partir de toute une collection de documents[5] de sorte que moins il y a de documents qui contiennent le mot, plus forte est sa spécificité. TF-IDF prenant en compte la totalité du corpus, il s'agit de montrer que TopicRank est capable d'extraire le contenu principal d'un document sans nécessiter de données supplémentaires et qu'il est donc applicable à des documents de tous types pour lesquels aucun travail de collecte préalable n'est nécessaire.

Au même titre que TopicRank, nous proposons une implémentation des méthodes de référence disponible sur la plate-forme de développement collaboratif GitHub[6]. Dans un souci de comparaison, lorsque les méthodes de référence partagent le même comportement que TopicRank, celui-ci est réalisé par le même composant.

4.1.2. *Collections de données*

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 dans ce travail diffèrent en termes de langue, nature, taille des documents et types d'annotateurs (auteurs, lecteurs ou les deux).

DUC (Over, 2001) est une collection en anglais issue des données de la campagne d'évaluation DUC-2001 qui 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). Nous utilisons ces 308 documents pour comparer TopicRank avec les méthodes de référence.

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 244 articles scientifiques (conférences et ateliers) issus de la bibliothèque numérique ACM. La collection est répartie en deux sous-ensembles : 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 pour comparer TopicRank avec les méthodes de référence et les 144 documents d'entraînement pour paramétrer TopicRank et SingleRank. En ce qui concerne les termes-clés associés aux documents, ce sont les termes-clés des auteurs combinés aux termes-clés donnés par des étudiants.

5. Dans ce travail, nous utilisons la collection dont est extrait le document.
6. `https://github.com/adrien-bougouin/KeyBench/tree/ijcnlp_2013`

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
Mots/document	900,7	5 177,7	308,5	6 839,4
Termes-clés/document	8,1	14,7	9,6	5,2
Mots/termes-clés	2,1	2,1	1,7	1,6
Termes-clés extractibles	96,5 %	77,9 %	92,4 %	78,9 %

Tableau 1. *Statistiques sur les données de test utilisées. Les termes-clés extractibles sont les termes-clés qui peuvent être extraits à partir du contenu des documents.*

WikiNews[7] est une collection de 100 articles journalistiques en français que nous avons extraits du site Web WikiNews[8] entre les mois de mai et de décembre 2012. Pour l'annotation en termes-clés, nous avons demandé à neuf étudiants de master en TAL d'extraire (librement) les termes-clés de 33 documents, de sorte que chaque document soit annoté par au moins trois étudiants. Nous utilisons les 100 documents de Wikinews pour comparer TopicRank avec les méthodes de référence.

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. Elle 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 pour comparer TopicRank avec les méthodes de référence et les 141 documents d'entraînement pour paramétrer TopicRank et SingleRank. Dans cette collection, seuls les termes-clés des auteurs sont disponibles.

Le tableau 1 donne les statistiques extraites des ensembles de test 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.

7. `https://github.com/adrien-bougouin/WikinewsKeyphraseCorpus`
8. `http://fr.wikinews.org`

4.1.3. *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 grammaticalement. La segmentation en mots est effectuée par le TreeBank-WordTokenizer, 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[9], pour le français. Quant à l'étiquetage grammatical, il est réalisé avec le Stanford POS tagger (Toutanova *et al.*, 2003) pour l'anglais et avec l'outil MElt (Denis et Sagot, 2009) pour le français. Tous ces outils sont utilisés avec leur configuration par défaut.

4.1.4. *Mesures d'évaluation*

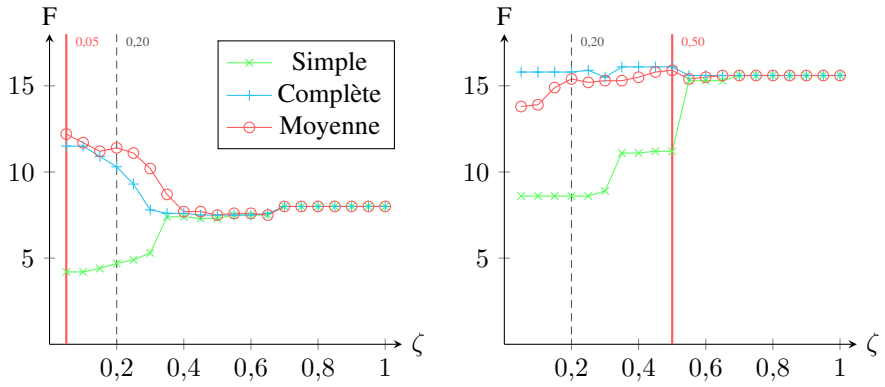
Les performances des méthodes d'extraction de termes-clés sont exprimées en termes de précision (P), rappel (R) et f-score (f1-mesure, F). En accord avec l'évaluation menée dans les travaux précédents, nous considérons correcte l'extraction d'une variante flexionnelle d'un terme-clé de référence (Kim *et al.*, 2010). Les opérations de comparaison entre les termes-clés de référence et les termes-clés extraits sont donc effectuées à partir de la racine des mots qui les composent en utilisant la méthode de racinisation de Porter (1980).

4.2. *Analyse empirique de TopicRank*

Dans cette section, nous effectuons des expériences préliminaires afin de déterminer quelle est la configuration optimale de TopicRank. En utilisant les ensembles d'entraînement de SemEval et de DEFT, nous réalisons deux expériences durant lesquelles nous faisons varier, dans un premier temps, le seuil de similarité (ζ) et la stratégie de groupement (simple, complète et moyenne), puis dans un second temps, la stratégie de sélection du terme-clé candidat le plus représentatif de chacun des sujets les plus importants.

La figure 4 présente les résultats de TopicRank lorsque nous faisons varier le seuil ζ avec un pas de 0,05 pour toutes les stratégies de groupement[10]. Globalement, chaque stratégie de groupement a un comportement qui lui est propre jusqu'à un certain point de convergence lorsque ζ vaut 0,70, ce point de convergence correspondant à la valeur du seuil ζ pour laquelle les sujets créés sont les mêmes quelle que soit la stratégie. Avec la stratégie simple, les résultats s'améliorent lorsque 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

9. `http://alpage.inria.fr/statgram/frdep/fr_stat_dep_parsing.html`
10. La stratégie de sélection du terme-clé le plus représentatif par sujet utilisée dans cette expérience est celle qui consiste à sélectionner le candidat qui apparaît en premier dans le document, pour chaque sujet.



(a) SemEval (b) DEFT

Figure 4. *Résultats de l'extraction de dix termes-clés avec TopicRank, en fonction de la stratégie de regroupement et de la valeur du seuil de similarité ζ, sur les ensembles d'entraînement de SemEval et de DEFT*

et la qualité du groupement s'améliore. En opposition, la stratégie complète, qui a le fonctionnement inverse, voit ses résultats se dégrader lorsque ζ augmente. Enfin, la stratégie moyenne agit en compromis. Pour SemEval, son comportement est le même que celui de la stratégie complète, mais ses résultats sont supérieurs jusqu'au point de convergence. Pour DEFT, son comportement est le même que celui de la stratégie simple, mais ses résultats sont très supérieurs jusqu'au point de convergence. Après observation des résultats de cette expérience, nous décidons d'utiliser la stratégie moyenne avec un seuil ζ de 0,20 pour toutes les expériences suivantes.

La figure 5 présente les résultats obtenus avec TopicRank et les différentes stratégies de sélection d'un terme-clé candidat par sujet. Les résultats confirment notre hypothèse qui est que le choix des candidats apparaissant en premier dans le document fournit de meilleurs termes-clés que le choix des candidats centroïdes ou des candidats les plus fréquents. La stratégie centroïde donne de très faibles résultats tandis que la stratégie fréquence n'est pas aussi stable que la stratégie position. Enfin, bien que la stratégie position donne les résultats les plus 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 de 14,8 points de f-score pour SemEval et de 5,4 points de f-score pour DEFT est encourageante pour de futurs travaux.

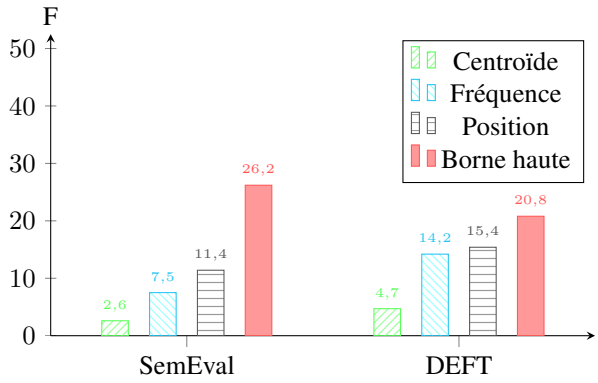


Figure 5. *Résultats de l'extraction de dix termes-clés, avec TopicRank, en fonction des différentes sélections de termes-clés candidats par sujet*

4.3. *Paramétrage empirique de SingleRank*

Contrairement aux autres méthodes de référence, SingleRank possède un paramètre qui est définit arbitrairement : la fenêtre de cooccurrences fixée à dix par Wan et Xiao (2008). De même que pour TopicRank, nous utilisons les ensembles d'entrainement de SemEval et de DEFT pour déterminer qu'elle est la valeur optimale de la fenêtre de cooccurrences pour SingleRank dans notre cadre expérimental[11].

La figure 6 présente les résultats de SingleRank lorsque nous faisons varier la fenêtre de cooccurrences de deux à vingt mots avec un pas de un. Globalement, nous observons une stabilité des performances de SingleRank quelle que soit la valeur utilisée pour la fenêtre de cooccurrences, avec des résultats optimaux obtenus lorsque celle-ci vaut 12. Dans les expériences suivantes, nous fixons donc ce paramètre à 12.

4.4. *Comparaison de TopicRank avec l'existant*

Le tableau 2 montre les performances de TopicRank comparées à celles des trois méthodes de référence. De manière générale, les performances des méthodes d'extraction de termes-clés sont basses. De plus, il est avéré que les documents de grande taille, tels que ceux de SemEval et de DEFT, sont plus difficiles à traiter que les autres documents. Ceci est dû au fait que, bien que les longs documents soient plus riches, le nombre de termes-clés candidats qui y sont sélectionnés est tellement important (par exemple environ 900 candidats sont sélectionnés par TopicRank pour chaque do-

11. Nous ne répétons pas cette expérience pour TextRank, car le critère d'adjacence (fenêtre de valeur 2) est un critère fort dans la méthode TextRank.

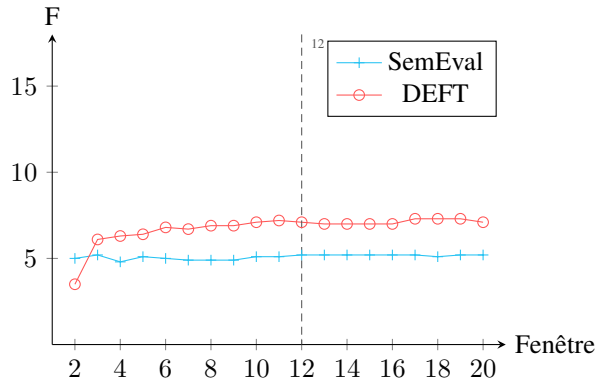


Figure 6. *Résultats de l'extraction de dix termes-clés, avec SingleRank, selon la fenêtre de cooccurrences utilisée*

cument de DEFT) que trouver les termes-clés parmi eux est plus difficile (Hasan et Ng, 2014).

Globalement, TopicRank donne de meilleurs résultats que les méthodes de référence utilisées. Comparé 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, alors que TopicRank n'utilise que le document à analyser. Comparé aux autres méthodes à base de graphe, TopicRank donne des résultats significativement meilleurs pour SemEval, WikiNews et DEFT. Ceci confirme donc que le groupement des candidats permet de rassembler des informations pour améliorer la précision de l'ordonnancement. En ce qui concerne DUC, notre méthode est aussi significativement meilleure que TextRank, mais elle ne l'est pas vis-à-vis de SingleRank. D'après la borne haute, l'une des raisons à la plus faible performance de TopicRank pour DUC est que la stratégie de sélection des candidats les plus représentatifs des sujets est moins adaptée. En effet, la différence avec la borne haute est de 12,9 points de f-score. Une analyse plus approfondie des différents apports de TopicRank peut aussi donner une piste sur les raisons de ses moins bons résultats.

Dans le but de confirmer la pertinence de tous les apports de TopicRank, nous réalisons une expérience supplémentaire dans laquelle nous appliquons individuellement à SingleRank toutes les modifications successives permettant d'obtenir la méthode TopicRank depuis la méthode SingleRank : l'usage d'un graphe complet (+ complet), la projection des termes-clés candidats dans le graphe (+ candidats) et la projection des sujets dans le graphe (+ sujets). Les résultats de ces trois variantes de SingleRank sont présentés dans le tableau 3. Globalement, l'usage des termes-clés candidats, ou sujets, induit une amélioration significative des performances de SingleRank, avec une amélioration plus importante en utilisant les sujets. Cela confirme la pertinence d'ordon-

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	22,6	28,8	25,0	4,8	3,3	3,9	19,2	20,4	19,5	4,7	9,4	6,2
TopicRank	18,2	23,2	20,1	**15,1**	**10,6**	**12,3**†	**34,8**	**37,3**	**35,4**†	**11,3**	**21,0**	**14,5**†
Borne haute	**31,6**	**35,3**	**33,0**	**33,8**	**23,3**	**27,3**	**41,7**	**44,1**	**42,2**	**14,5**	**27,0**	**18,7**

Tableau 2. *Résultats de l'extraction de dix 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	22,6	28,8	25,0	4,8	3,3	3,9	19,2	20,4	19,5	4,7	9,4	6,2
+ 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	10,4	13,5	11,6	9,4	6,8	7,8†	28,5	30,0	28,8†	10,3	19,2	13,2†
+ sujets	18,9	24,2	21,0	14,2	9,9	11,6†	30,7	32,6	31,1†	11,1	20,4	14,2†
TopicRank	18,2	23,2	20,1	**15,1**	**10,6**	**12,3**†	**34,8**	**37,3**	**35,4**†	**11,3**	**21,0**	**14,5**†

Tableau 3. *Résultats de l'extraction de dix 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.*

ner directement les candidats, plutôt que les mots, ainsi que la pertinence de grouper les candidats représentant le même sujet afin de mutualiser les 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 cooccurrences. Toutefois, nous pensons que l'usage du graphe complet est à privilégier afin d'éviter d'avoir à fixer le paramètre de la fenêtre de cooccurrences.

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 cooccurrences que lorsque les nœuds sont les mots du document et est donc moins précis pour l'ordonnancement.

5. Analyse d'erreurs

Dans cette section, nous proposons d'analyser les erreurs de TopicRank. Dans un premier temps, nous analysons les sujets que détecte TopicRank, puis dans un second

temps, nous analysons les termes-clés de référence qui ne sont pas extraits par TopicRank.

5.1. *Analyse des sujets détectés*

Dans cette section, nous analysons les groupements en sujets effectués par TopicRank afin de déterminer quelles sont les principales causes d'erreurs.

Nous observons des erreurs liées à la sélection des termes-clés candidats. Lors de cette étape, certaines unités textuelles sont sélectionnées comme candidats à cause d'erreurs commises lors de l'étiquetage grammatical. Ces erreurs concernent principalement la détection des participes. Par exemple, dans la phrase « [...] elles ne cessent de se développer à travers le monde et particulièrement dans les pays dits "du sud" [...] »[12], « dits » est un adjectif selon l'outils MElt, ce qui entraîne la sélection erronée du terme-clé candidat « pays dits ».

Nous observons également de nombreuses erreurs lorsque les groupements sont déclenchés par un adjectif. Ce sont particulièrement les expansions nominales s'effectuant à gauche qui en sont la cause (par exemple « même langue » groupé avec « même représentation »). 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 sont liés au contexte général du document, ils sont très fréquemment utilisés et beaucoup de candidats les contenant sont groupés par erreur (par exemple « forces économiques » peut être groupé avec « délabrement économique » dans un document d'économie). Outres ces groupements erronés, nous observons aussi de mauvais groupements lorsque les candidats ne contiennent que très peu de mots. 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.

5.2. *Analyse des faux négatifs*

Dans cette section, nous analysons les termes-clés de référence qui n'ont pas été extraits par TopicRank. Plus particulièrement, nous nous intéressons à ceux qui sont présents dans les dix sujets jugés les plus importants de chaque document, mais qui n'ont pas été sélectionnés pour les représenter. Nous observons deux sources d'erreurs.

La première source d'erreurs est le groupement en sujets. Lorsqu'un sujet détecté contient en réalité des termes-clés candidats représentant des sujets différents, la stratégie de sélection du meilleur terme-clé dans le sujet parvient à sélectionner le terme-clé correct dans certains cas, mais elle échoue parfois.

12. Exemple issu de l'article d'anthropologie *Le marché parallèle du médicament en milieu rural au Sénégal* (`http://id.erudit.org/iderudit/014935ar`) de la collection DEFT.

La seconde source d'erreurs est la spécialisation des termes-clés de référence. Nous observons deux problèmes de sous et sur-spécialisation de certains termes-clés extraits vis-à-vis des termes-clés de référence. Dans le cas de la sous-spécialisation, nous pouvons citer, par exemple, « papillons » qui est extrait à la place de « papillons mutants »[13]. Bien que ce problème de sous-spécialisation soit identifié, l'existance du problème inverse le rend plus difficile à résoudre. Dans le cas de la sur-spécialisation, nous pouvons citer, par exemple, « député Antoni Pastor » qui est extrait à la place de « Antoni Pastor »[14]. La raison principale de ce problème est l'aspect libre de l'annotation manuelle des termes-clés. Toutefois, privilégier les modifications adjectivales (par exemple « mutants ») et, au contraire, éviter les modifications nominales (par exemple « député ») semblent être une hypothèse à vérifier.

6. 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 en 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. Le groupement des termes-clés potentiels en sujets distincts permet de rassembler des indices utiles auparavant éparpillés et le choix d'un seul terme-clé pour représenter un sujet important permet d'extraire un ensemble de termes-clés non redondants (pour k termes-clés extraits, exactement k sujets sont couverts). Enfin, le graphe est complet et ne requiert plus le paramétrage d'une fenêtre de cooccurrences, contrairement aux autres méthodes à base de graphe.

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 qu'il est encore possible d'améliorer notre méthode en proposant une nouvelle stratégie de sélection du terme-clé candidat le plus représentatif d'un sujet (pour un gain maximum allant de 4,2 à 15 points de f-score).

Nous avons aussi effectué une analyse d'erreurs à partir de laquelle trois perspectives de travaux futurs émergent.

Nous avons pour objectif d'améliorer la sélection des termes-clés candidats. Aussi, des méthodes empruntées à d'autres domaines du TAL peuvent être appliquées. Il semble, par exemple, pertinent d'évaluer l'apport des méthodes d'extraction terminologiques (Castellví *et al.*, 2001) pour la sélection des termes-clés candidats.

13. Exemple issue de l'article journalistique *Fukushima fait muter les papillons* (`http://fr.wikinews.org/w/index.php?oldid=432477`) de la collection WikiNews.
14. Exemple issu de l'article journalistique *Îles Baléares : le Parti populaire exclut le député Antoni Pastor pour avoir défendu la langue catalane* (`http://fr.wikinews.org/w/index.php?oldid=479948`) de la collection WikiNews.

Nous envisageons également d'améliorer le groupement en sujets, car celui-ci est très naïf et ne tient compte ni de la synonymie, ni de l'ambiguïté des mots. De plus, l'usage du radical (Porter, 1980) des mots n'est pas sans introduire du bruit lié à certains faux positifs. L'ajout de connaissances concernant les synonymes permettrait de créer des sujets plus complets et une étape de désambiguïsation éviterait un groupement systématique des termes-clés candidats ayant un ou plusieurs mots en commun. Nous envisageons aussi de remplacer la racinisation par une méthode de lemmatisation. D'un point de vue plus technique, il faudrait explorer différentes méthodes 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.

Enfin, une étude détaillée des caractéristiques des termes-clés pourrait orienter notre travail vers des critères plus efficaces pour la définition d'une stratégie « optimale » de sélection du terme-clé le plus représentatif d'un sujet. Un apprentissage supervisé à partir de certains critères est aussi envisagé, 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).

7. 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.

Bougouin A., Boudin F., Daille B., « TopicRank : Graph-Based Topic Ranking for Keyphrase Extraction », *Proceedings of the 6th International Joint Conference on Natural Language Processing (IJCNLP)*, Asian Federation of Natural Language Processing, Nagoya, Japan, p. 543-551, October, 2013.

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.

Castellví M. T. C., Bagot R. E., Palatresi J. V., « Automatic Term Detection : A Review of Current Systems », *Recent Advances in Computational Terminology*, vol. 2, p. 53-88, 2001.

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 (IJCNLP)*, 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 (SemEval)*, 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 (COLING)*, Association for Computational Linguistics, Stroudsburg, PA, USA, p. 365-373, 2010.

Hasan K. S., Ng V., « Automatic Keyphrase Extraction : A Survey of the State of the Art », *Proceedings of the Association for Computational Linguistics (ACL)*, Association for Computational Linguistics, Baltimore, Maryland, June, 2014.

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

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 (SemEval)*, 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(EMNLP)*, 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 (EMNLP)*, 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 (EMNLP)*, 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 (AAAI)*, 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 (SemEval)*, Association for Computational Linguistics, Stroudsburg, PA, USA, p. 162-165, 2010.

Pearce D., « A Comparative Evaluation of Collocation Extraction Techniques », *In third International Conference on Language Resources and Evaluation (LREC)*, Citeseer, 2002.

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 : Human Language Technology - Volume 1 (NAACL)*, 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.

Wang R., Liu W., McDonald C., « How Preprocessing Affects Unsupervised Keyphrase Extraction », *in* A. Gelbukh (ed.), *Computational Linguistics and Intelligent Text Processing*, vol. 8403 of *Lecture Notes in Computer Science*, Springer Berlin Heidelberg, p. 163-176, 2014.

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.