

## Résumé automatique multi-document et indépendance de la langue : une première évaluation en français

Florian Boudin<sup>1</sup> Juan-Manuel Torres-Moreno<sup>1, 2</sup>

(1) LIA / Université d'Avignon, 339 chemin des Meinajariès, 84911 Avignon

(2) École Polytechnique de Montréal, CP 6079 Succ. Centre-ville, Montréal

{florian.boudin, juan-manuel.torres}@univ-avignon.fr

**Résumé.** Le résumé automatique de texte est une problématique difficile, fortement dépendante de la langue et qui peut nécessiter un ensemble de données d'apprentissage conséquent. L'approche par extraction peut aider à surmonter ces difficultés. (Mihalcea, 2004) a démontré l'intérêt des approches à base de graphes pour l'extraction de segments de texte importants. Dans cette étude, nous décrivons une approche indépendante de la langue pour la problématique du résumé automatique multi-documents. L'originalité de notre méthode repose sur l'utilisation d'une mesure de similarité permettant le rapprochement de segments morphologiquement proches. De plus, c'est à notre connaissance la première fois que l'évaluation d'une approche de résumé automatique multi-document est conduite sur des textes en français.

**Abstract.** Automatic text summarization is a difficult task, highly language-dependent and which may require a large training dataset. Recently, (Mihalcea, 2004) has shown that graph-based approaches applied to the sentence extraction issue can achieve good results. In this paper, we describe a language-independent approach for automatic multi-document text summarization. The main originality of our approach is the use of an hybrid similarity measure during the graph building process that can identify morphologically similar words. Moreover, this is as far as we know, the first time that the evaluation of a summarization approach is conducted on French documents.

**Mots-clés :** Résumé automatique de texte, Approches à base de graphes, Extraction d'information.

**Keywords:** Text summarization, Graph-Based approaches, Information Extraction.

## 1 Introduction

Un résumé est un texte reformulé dans un espace plus réduit. Il doit exprimer avec un minimum de mots le contenu essentiel d'un document. Son but est d'aider le lecteur à repérer les informations qui peuvent l'intéresser sans pour autant devoir lire le document en entier. Mais pourquoi avons-nous tant besoin de résumés ? Simplement parce que nous ne disposons pas d'assez de temps et d'énergie pour tout lire. La masse d'information textuelle sous forme électronique ne cesse d'augmenter, que ce soit sur Internet ou dans les réseaux des entreprises. Ce volume croissant de textes disponibles rend difficile l'accès à l'information désirée sans l'aide d'outils spécifiques. Mais produire un résumé est une tâche très complexe car elle nécessite des connaissances linguistiques ainsi que des connaissances du monde qui restent très difficiles à

incorporer dans un système automatique. Il existe néanmoins des approches permettant d'imiter une partie du processus cognitif du résumé. Ces dernières peuvent être regroupées en deux catégories, **les méthodes extractives et abstractives**. Bien que des méthodes par abstraction ont été mises au point, les outils nécessaires à la compréhension sémantique du texte ou à la génération de texte en langue naturelle n'ont pas atteint la maturité indispensable à une utilisation robuste. L'approche que nous proposons repose sur un processus extractif. Dans ce paradigme, les segments textuels –le plus souvent des phrases– contenant les idées essentielles du document sont extraits puis assemblés afin de produire un résumé, également appelé extrait (Mani & Maybury, 1999).

Il existe de nombreuses variantes de résumé automatique. La plus simple étant le résumé générique mono-document où il s'agit de produire un résumé en préservant au mieux toutes les idées essentielles contenues dans un document. Cette variante, qui pourtant paraît être la plus simple, **pose encore de nombreux problèmes**. En effet, du type de document que l'on veut résumer dépend les performances des systèmes. Il est envisageable de générer un résumé à partir d'un article de journal tandis qu'il est, du moins actuellement, **quasiment impossible de le faire à partir d'œuvres littéraires** (Mihalcea & Ceylan, 2007). Par opposition au résumé générique, la tâche de résumé orienté consiste en la production d'un résumé qui satisfait les besoins d'un utilisateur. Ces besoins, généralement exprimés au moyen d'une requête, doivent permettre au système d'isoler les parties du document concernant une (plusieurs) thématique(s) précise(s) pour ensuite produire un résumé n'incluant que ces dernières. À ces deux variantes, peut s'ajouter la problématique des résumés **multi-documents**. **De nouvelles difficultés sont alors introduites** : les phrases extraites à partir de documents différents peuvent être complémentaires, redondantes ou contradictoires. Il faudra donc veiller à la cohésion des phrases mais surtout à la cohérence du résumé (absence de contradiction ou de redondance dans l'enchaînement des phrases).

Évaluer la qualité d'un résumé est un problème difficile auquel la communauté n'a pour le moment su répondre qu'avec des solutions partielles. En effet, il n'existe pas de résumé « idéal ». **Les résumés écrits par des personnes différentes ne sont pas toujours convergents au niveau du contenu**. La rédaction de ce type de document requiert une analyse du texte afin d'en dégager les idées, le style et les arguments, ce que chaque personne fait de manière différente. Par conséquent, deux résumés de contenu informationnel très similaire peuvent être produits en utilisant un vocabulaire totalement différent. De manière générale, les méthodes d'évaluation peuvent être classées en deux catégories (Spärck Jones & Galliers, 1996). La première regroupe **les évaluations dites extrinsèques**, les résumés sont évalués en se basant sur leur aptitude à accélérer la complétion de tâches annexes (e.g. l'utilisation des résumés, à la place des documents sources, dans des systèmes question/réponse ou de classification de documents). La deuxième catégorie réunit **les évaluations intrinsèques**, où les résumés sont alors jugés directement en se basant sur leur analyse. Cette tâche peut être réalisée manuellement (des juges évaluant les qualités d'un résumé comme la lisibilité, la complexité de la langue ou la présence des concepts majeurs du document source) ou automatiquement en calculant des mesures de similarité entre le résumé candidat et un ou plusieurs résumés de référence. **Ce n'est que très récemment que la communauté a su mettre à disposition des mesures d'évaluation automatique pertinentes** (Lin, 2004) ainsi que des ensembles de données de qualité (Nenkova, 2005). Le problème majeur restant que la langue utilisée dans ces corpora est presque toujours l'anglais. (Châar *et al.*, 2004) ont néanmoins proposé un protocole d'évaluation sur des documents en français qui utilise les données des campagnes d'évaluation CLEF. Bien que ne portant pas sur la qualité linguistique des résumés, les mesures d'évaluation reportées montrent le degré de précision avec lequel les unités informatives peuvent être extraites par leur approche.

Dans cet article, nous proposons d'étudier le comportement de plusieurs approches extractives indépendantes de la langue des documents à traiter. La problématique que nous abordons est le résumé générique multi-documents d'articles de journaux en français à partir de méthodes à base de graphes. Nous présentons en section 2 un rapide état de l'art sur le résumé automatique de texte et la problématique d'adaptabilité. Nous montrons en section 3 la méthode de construction du graphe et la sélection des segments basée sur son analyse. Nous évaluons notre approche en section 4 en la comparant aux résumés de référence produits manuellement. Nous discutons des résultats obtenus en section 5, et proposons des perspectives de recherche.

## 2 Travaux précédents

(Luhn, 1958) a probablement été le premier à proposer une méthode numérique pour la production d'extraits. Déjà motivé par la problématique de surcharge d'information face à des quantités qui peuvent paraître dérisoires presque 50 ans plus tard, Luhn décrit une technique simple, spécifique aux articles scientifiques qui utilise la distribution des fréquences de mots dans le document pour pondérer les phrases. Par la suite, (Edmunson, 1969) est allé plus loin en introduisant des critères comme la position des phrases, la présence de mots provenant de la structure du document (i.e. titres, sous-titres, etc.) ou la présence de mots indices (*cue words*). Leurs travaux ont eu un impact considérable, la grande majorité des approches d'aujourd'hui étant toujours basées sur ces mêmes idées. Quelques-unes de ces approches, parmi lesquelles certaines font partie des plus performantes (Teufel & Moens, 1997; Hirao *et al.*, 2002), utilisent des algorithmes supervisés qui tentent de caractériser ce qui fait un bon résumé. Cependant, le prix à payer pour obtenir de bons résultats est de disposer d'un ensemble de données d'apprentissage. C'est ce point qui rend ce type d'approche difficilement adaptable à d'autres langues ou domaines.

Il existe néanmoins quelques méthodes destinées à résoudre ce problème. Ainsi, (Mihalcea, 2004; Erkan & Radev, 2004) proposent de considérer le processus extractif comme une identification des segments les plus prestigieux dans un graphe. Les algorithmes de classement basés sur les graphes tel que *PageRank* (Brin & Page, 1998) ont été utilisés avec succès dans les réseaux sociaux, l'analyse du nombre de citations ou l'étude de la structure du Web. Ces algorithmes peuvent être vus comme les éléments clés du paradigme amorcé dans le domaine de la recherche sur Internet, à savoir le classement des pages Web par l'analyse de leurs positions dans le réseau et non pas de leurs contenus. En d'autres termes, ces algorithmes permettent de décider de l'importance du sommet d'un graphe en se basant non pas sur l'analyse locale du sommet lui-même, mais sur l'information globale issue de l'analyse récursive du graphe complet. Appliqué au résumé automatique cela signifie que le document est représenté par un graphe d'unités textuelles (phrases) liées entre elles par des relations issues de calculs de similarité. Les phrases sont ensuite sélectionnées selon des critères de centralité ou de prestige dans le graphe puis assemblées pour produire des extraits. Les résultats reportés montrent que les performances des approches à base de graphe sont au niveau des meilleurs systèmes actuels (Mihalcea, 2005) mais ne portent que sur des documents en anglais et en portugais. Deux questions se posent alors : que doit-on attendre des résultats sur des documents en français ? peut-on améliorer les performances des systèmes à base de graphes ?

Il est important de noter que les méthodes de classement sont entièrement dépendantes de la bonne construction du graphe sensé représenter le document. Puisque ce graphe est généré à partir de mesures de similarités inter-phrases, l'impact que peut avoir le choix de la méthode

de calcul est à considérer. Dans leurs travaux, (Mihalcea, 2004; Erkan & Radev, 2004) utilisent le modèle en sac-de-mots pour représenter chaque phrase comme un vecteur à  $N$  dimensions, où  $N$  est le nombre total de mots différents du regroupement et chaque composante du vecteur un poids  $tf \times idf$ . Les valeurs de similarité entre phrases sont ensuite obtenues par un calcul du cosinus entre leurs représentations vectorielles. Le point faible de cette mesure, et plus généralement de toutes les mesures utilisant les mots comme unités, est qu'elles sont tributaires du vocabulaire. Dans une optique d'indépendance de la langue, les pré-traitements qui sont appliqués aux segments se doivent d'être minimaux. C'est malheureusement dans cette configuration que les performances de la mesure cosinus chutent car elle ne permet en aucun cas de mettre en relation des mots qui morphologiquement peuvent être très proches. Une solution peut venir de la combinaison avec des mesures de similarité basées sur les caractères. (Boudin *et al.*, 2008) proposent une mesure dérivée d'un calcul de similarité entre chaînes de caractères originellement employé pour la détection d'entités redondantes (*Record Linkage*). Cette mesure permet de créer des relations entre deux segments qui même s'il ne partagent aucun mot, en contiennent des morphologiquement proches. Une seconde question est donc de savoir si la construction du graphe du document à partir de mesures mixtes (mots et caractères) permet d'améliorer l'extraction de segments.

## 3 Méthode

### 3.1 Construction du graphe

Afin de permettre aux algorithmes de classement d'être appliqués sur des documents en langage naturel, nous devons construire un graphe qui représente le texte et interconnecte les unités textuelles avec des relations de sens. La nature des relations et la taille des unités dépend bien entendu du type d'application que l'on cible. Pour le résumé automatique le choix le plus simple au niveau de la taille des unités est la phrase complète. En effet, cela permet lors de la génération du résumé de ne pas avoir à se soucier de la grammaticalité des segments assemblés. Une connexion entre deux phrases est définie comme une mesure de similarité morphologique. Ce type de relation peut être vu comme une *recommandation*. Une phrase qui traite de certains concepts du texte donne au lecteur une *recommandation* quant aux autres phrases partageant du contenu en commun.

Les segments ont tout d'abord été nettoyés de leurs ponctuation et la casse normalisée<sup>1</sup>. Le seul pré-traitement statistique qui leur est appliqué correspond à un filtrage des mots communs à l'aide d'une liste<sup>2</sup>. Il s'agit d'un processus simple et facilement adaptable à d'autres langues ou domaines. Pour chaque ensemble de documents, un graphe non dirigé et valué  $G = (S, A)$  est construit.  $S$  est l'ensemble de sommets du graphe et  $A$  est l'ensemble des arêtes,  $A \subseteq S \times S$ . La valeur de chaque arête est évaluée à l'aide d'une mesure de similarité calculée entre les deux nœuds de la connexion. Dans les formulations originelles de (Mihalcea, 2004; Erkan & Radev, 2004), la mesure de similarité utilisée est le cosinus. Nous proposons de la combiner avec une mesure morphologique de plus longue sous-chaîne de caractères, qui nous pensons, peut permettre d'augmenter la qualité des connexions entre segments (voir équation 1).

1. Transformation des majuscules en minuscules.

2. Les mots outils et non informatifs sont supprimés des segments.

$$sim(S_1, S_2) = \alpha \cdot \cos(S_1, S_2) + (1 - \alpha) \cdot \text{LCS}^*(S_1, S_2); \quad 0 < \alpha < 1 \quad (1)$$

Le paramètre  $\alpha$  a été fixé empiriquement à 0,9.

L'équation 2 montre la mesure LCS (*Longest Common Substring*) que nous avons modifiée en  $\text{LCS}^*$  pour calculer la similarité entre deux segments  $S_1$  et  $S_2$ .

$$\text{LCS}^*(S_1, S_2) = \frac{1}{|S_1|} \cdot \sum_{m_1 \in S_1} \max_{m_2 \in S_2'} \text{LCS}(m_1, m_2) \quad (2)$$

où  $S_2'$  est l'ensemble de mots du segment  $S_2$  dans lequel les mots  $m_2$ , qui ont déjà maximisé  $\text{LCS}(m_1, m_2)$  durant les étapes précédentes du calcul, ont été enlevés. Le facteur  $|S_1|^{-1}$  permet de normaliser le calcul.

Pour réduire le bruit qui peut être introduit par une mesure calculée sur les caractères, nous avons ajouté un seuil qui filtre les valeurs minimales. La somme est faite uniquement pour des valeurs de maximum supérieur à 0,6, c'est à dire pour deux mots partageant au moins 60% des caractères en commun. Ce traitement, qui remplace avantageusement une lemmatisation ou *stemming* classique, rend notre méthode plus indépendante de la langue. Un exemple de graphe est montré dans la Table 2 (section annexe).

## 3.2 Pondération des segments

Le paradigme extractif pour le résumé automatique a permis la mise au point de méthodes performantes qui peuvent se passer d'un ensemble de données d'apprentissage. Le document est représenté par un graphe d'unités textuelles (les phrases) liées entre elles par des relations issues de mesures de similarité. Les algorithmes de classement permettent ensuite de décider de l'importance d'un segment en se basant sur l'information globale issue de l'analyse récursive du graphe. Les méthodes extractives fonctionnent par une sélection d'un sous-ensemble de phrases. Ce processus peut être vu comme une identification des segments les plus centraux, *i.e.* contenant les idées essentielles du texte original. Cette section décrit les algorithmes de classement que nous avons utilisé pour extraire cet sous-ensemble de phrases à partir d'un graphe.

### 3.2.1 Popularité

Cette méthode est une interprétation naïve du phénomène de popularité : une phrase est considérée importante si elle est reliée à un grand nombre d'autres phrases dans le graphe. Le score de popularité de chaque sommet est calculé à partir du nombre d'arêtes entrantes.

$$\text{popularité}(s) = \text{card}\{\text{adj}[s]\} \quad (3)$$

où  $\text{card}\{\text{adj}[s]\}$  est la cardinalité de l'ensemble de sommets reliés à  $s$  dans une matrice d'adjacence. Il est possible d'utiliser un seuil (empirique) afin d'éliminer certaines arêtes que l'on peut juger comme peu significatives, car leurs valeurs sont très faibles.

### 3.2.2 LexRank

*PageRank* (Brin & Page, 1998) est sans doute le plus populaire des algorithmes de classement, conçu à l'origine pour déterminer l'importance d'une page Web. (Erkan & Radev, 2004) proposent une interprétation de cet algorithme, dénommée *LexRank*, pour le classement des segments textuels. Contrairement à la méthode originelle de *PageRank*, le graphe de segments est construit à partir de mesures de similarité symétriques, il est par conséquent non dirigé. Le score de chaque sommet  $s$  est calculé itérativement jusqu'à la convergence<sup>3</sup> par :

$$p(s) = (1 - d) + d \times \sum_{v \in \text{adj}[s]} \frac{p(v)}{\text{popularité}(v)} \quad (4)$$

où  $\text{popularité}(v)$  est le nombre d'arêtes du sommet  $v$  et  $d$  est un facteur d'amortissement (*damping factor*) généralement fixé à 0,85.

### 3.2.3 TextRank

*TextRank* est une variante de l'algorithme *LexRank* dans laquelle les valeurs de similarité assignées aux arcs sont utilisées pour la pondération des sommets. De cette manière l'impact des sommets connectés par des arcs de valeurs faibles est minimisé dans le calcul du score du segment. Le score de chaque sommet  $s$  est calculé itérativement jusqu'à la convergence par :

$$p(s) = (1 - d) + d \times \sum_{v \in \text{adj}[s]} \frac{\text{Sim}(s, v)}{\sum_{z \in \text{adj}[v]} \text{Sim}(z, v)} p(v) \quad (5)$$

## 3.3 Assemblage des résumés

Une fois les phrases pondérées, elles doivent être sélectionnées et assemblées afin de produire le résumé. La taille des résumés est fixée par un nombre de mots maximum. L'algorithme de production que nous proposons tente de maximiser le nombre de mots dans le résumé tout en privilégiant les segments de plus hauts scores. La redondance intra-résumé est minimisée par un simple seuil de similarité inter-phrases appliqué en amont. De nombreuses contraintes s'ajoutent lors de la concaténation des segments. En effet, les segments doivent être ordonnés temporairement dans le résumé, c'est à dire en respectant les dates de publications des documents sources mais aussi la position dans le document si deux segments proviennent de la même source. De plus, un ensemble de post-traitements est appliqué aux segments qui dans la plupart des cas modifie leurs tailles. Il convient donc de produire le résumé en plusieurs passes, l'assemblage des segments devant être dynamiquement modifiable en fonction des contraintes. Les post-traitements suivants sont effectués en considérant l'ordre d'apparition des phrases dans le résumé : *i*) suppression du contenu entre parenthèses, *ii*) normalisation des références temporelles<sup>4</sup> et *iii*) normalisation de la ponctuation.

3. Un seuil d'acceptation d'erreur  $e = 0.0001$  permet d'arrêter les itérations lorsque les valeurs de tous les sommets n'ont pas été modifiés de plus de  $e$ .

4. Les dates complètes de la forme « 15 décembre 1982 » sont remplacées par « 15/12/1982 ».



## 4 Évaluation

Depuis 2001, le *National Institute of Standards and Technology*<sup>5</sup> (NIST) organise la campagne d'évaluation *Document Understanding Conference* (DUC), devenue depuis 2008 la campagne *Text Analysis Conference* (TAC)<sup>6</sup>. Son but est de promouvoir les progrès réalisés dans le domaine du résumé automatique de textes mais surtout de permettre aux chercheurs de participer à des expérimentations de grande envergure tant au point de vue du développement que de l'évaluation de leurs systèmes. Dans le cadre de ces campagnes, l'évaluation des systèmes est réalisée de manière intrinsèque sur le fond ainsi que sur la forme des résumés produits. Plusieurs notations sont attribuées manuellement aux résumés. Elle sont complétées par le calcul d'un ensemble de mesures semi-automatiques qui, au travers de mesures de similarités calculées entre un résumé candidat et un ou plusieurs résumés de référence, permettent de juger de la qualité du contenu. C'est ce type de mesures que nous avons utilisé pour évaluer la qualité de nos résumés.

L'évaluation de nos méthodes a été conduite en suivant un protocole très similaire à celui du NIST. Un corpus composé de 20 thématiques différentes (par exemple la « Visite du Dalai Lama en France », « Les Jeux Olympiques à Pekin », etc.) a été constitué<sup>7</sup>. Pour chacune de ses thématiques, un regroupement (*cluster*) de 10 articles de journaux de sources différentes a été assemblé. Quatre annotateurs ont ensuite produit manuellement quatre résumés de référence pour chaque *cluster*. Les recommandations principales qui leur ont été données étaient de créer un résumé d'un maximum de 100 mots, contenant les idées essentielles de l'ensemble de documents tout en utilisant un minimum de connaissances externes. La tâche du système de résumé automatique est de produire un résumé d'un maximum de 100 mots à partir de chaque *cluster* de documents. Nous utilisons les mesures semi-automatiques ROUGE (*Recall-Oriented Understudy for Gisting Evaluation*) (Lin, 2004) calculées à partir des quatre résumés de référence pour attribuer un score à chacune des méthodes que nous voulons évaluer.

Nous avons comparé les scores obtenus avec les méthodes par classement de *popularité*, *Lex-Rank* et *TextRank* sur des graphes construits à partir de mesures de similarité cosinus et de mesures de similarité mixtes (équation 1). À titre de comparaison, deux autres méthodes ont été ajoutées. La première est une *baseline* bien connue qui consiste à produire un résumé à partir des premières phrases du document le plus récent, et qui est d'ailleurs très difficile à battre car elle garde à la fois, l'information essentielle et la cohérence. La seconde méthode a été proposée par (Radev *et al.*, 2004) et suggère de considérer le processus extractif comme une identification des segments les plus centraux du *cluster*. La centralité d'un segment est définie en fonction de la centralité des mots qu'il contient. Une manière simple d'évaluer la centralité est alors de construire le centroïde, ce dernier pouvant être vu comme un pseudo-document composé des mots ayant un poids  $tf \times idf$  supérieur à un seuil prédéfini. Les segments partageant le plus de mots avec le centroïde sont considérés comme centraux et seront assemblés pour générer le résumé. Cette méthodologie a conduit au développement du premier système de résumé multi-documents sur le Web (Radev *et al.*, 2001).

Les résultats de cette évaluation sont montrés dans la table 1. On peut constater que les scores obtenus sur des graphes construits avec des mesures de similarité mixtes sont toujours meilleurs que ceux obtenus avec cosinus. Il est également intéressant de noter que toutes les méthodes

5. <http://www.nist.gov>

6. <http://www.nist.gov/tac>

7. Le corpus sera disponible sur le site du projet RPM2.

à base de graphes ont de meilleurs scores que la *baseline*. À titre de comparaison, le score ROUGE-1 moyen obtenu par le système *LexRank* (Erkan & Radev, 2004) (meilleur système sur la tâche de résumé automatique de la campagne DUC 2004) était de 0,396.

Evaluation	Popularité		LexRank		TextRank		Centroïde	Baseline
	<i>cosinus</i>	<i>mixte</i>	<i>cosinus</i>	<i>mixte</i>	<i>cosinus</i>	<i>mixte</i>		
ROUGE-1	0,37634	0,38702	0,36943	0,38934	0,36842	0,38994	0,31236	0,34661
ROUGE-2	0,11228	0,11539	0,10587	0,11625	0,10301	0,12318	0,06952	0,08324
ROUGE-SU4	0,13759	0,13993	0,13320	0,14126	0,13273	0,14836	0,09876	0,11470

TABLE 1 – Résultats des méthodes de pondération.

## 5 Conclusion et discussion

Nous avons présenté une première évaluation formelle pour la tâche du résumé automatique multi-documents en français. La méthode que nous avons proposée est basée sur une approche à base de graphes et se veut être la plus indépendante de la langue possible. Des tests sur un corpus de textes en français ont été réalisés et évalués selon un protocole similaire à celui utilisé par le NIST. Nous avons introduit une mesure mixte combinant le cosinus et la somme des distances morphologiques entre mots, ce qui a permis d'améliorer les résultats pour toutes les méthodes de classement. La détection numérique de la similarité morphologique évite les traitements de lemmatisation qui peuvent s'avérer être très lourds et trop dépendants de la langue. La tâche de résumé multi-documents est beaucoup plus complexe que celle du résumé mono-document, et les phénomènes de cohésion et de cohérence y sont bien plus gênants. Afin d'améliorer la qualité du résumé, plusieurs stratégies ont vu le jour. Bien que les graphes dirigés (*backward* et *forward*) semblent améliorer les résultats en mono-document (Mihalcea, 2004), en multi-document la problématique est différente. La prise en compte de la contrainte de temporalité des phrases ne garantit pas la qualité de la cohérence dans le résumé. Des recherches plus approfondies doivent être entreprises dans cette voie.

## Remerciements

Ce travail à été financé par le projet ANR RPM2 (Résumé Plurimédia, Multi-documents et Multi-opinions), <http://labs.sinequa.com/rpm2>. Les auteurs remercient Claude de Loupy, Christelle Ayache et Somara Seng pour avoir rendu l'évaluation de notre approche possible.

## Références

- BOUDIN F., TORRES-MORENO J.-M. & VELÁZQUEZ-MORALES P. (2008). An Efficient Statistical Approach for Automatic Organic Chemistry Summarization. In B. NORDSTRÖM & A. RANTA, Eds., *6th International Conference on Natural Language Processing, GoTAL 2008*, volume 5221 of *Lecture Notes in Computer Science*, p. 89–99, Gothenburg, Sweden : Springer.
- BRIN S. & PAGE L. (1998). The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems*, **30**(1-7), 107–117.



CHÂAR S., FERRET O. & FLUHR C. (2004). Filtrage pour la construction de résumés multi-documents guidée par un profil. *Traitement automatique des langues*, **45**(1), 65–93.

EDMUNSON H. P. (1969). New Methods in Automatic Extracting. *Journal of the Association for Computing : Machinery*, **16**(2), 264–285.

ERKAN G. & RADEV D. R. (2004). LexRank : Graph-based Lexical Centrality as Saliency in Text Summarization. *Journal of Artificial Intelligence Research*, **22**(2004), 457–479.

HIRAO T., SASAKI Y., ISOZAKI H. & MAEDA E. (2002). NTT's text summarization system for duc-2002. In *Document Understanding Conference (DUC)*, p. 104–107, Philadelphia, Pennsylvania, USA.

LIN C.-Y. (2004). Rouge : A package for automatic evaluation of summaries. In *Text Summarization Branches Out : Proceedings of the ACL-04 Workshop*, p. 74–81, Barcelona, Spain : Association for Computational Linguistics.

LUHN H. P. (1958). The automatic creation of literature abstracts. *IBM Journal of Research and Development*, **2**(2), 159–165.

MANI I. & MAYBURY M. T. (1999). *Advances in Automatic Text Summarization*. MIT Press.

MIHALCEA R. (2004). Graph-based ranking algorithms for sentence extraction, applied to text summarization. In *ACL 2004 on Interactive poster and demonstration sessions*, p. 181–184 : Association for Computational Linguistics Morristown, NJ, USA.

MIHALCEA R. (2005). Language independent extractive summarization. In *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, p. 49–52, Ann Arbor, Michigan : Association for Computational Linguistics.

MIHALCEA R. & CEYLAN H. (2007). Explorations in Automatic Book Summarization. In *Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, p. 380–389.

NENKOVA A. (2005). Automatic Text Summarization of Newswire : Lessons Learned from the Document Understanding Conference. In *National Conference on Artificial Intelligence (AAAI'05)*, Pittsburgh, Pennsylvania, USA.

RADEV D. R., BLAIR-GOLDENSOHN S. & ZHANG Z. (2001). Experiments in single and multi-document summarization using MEAD. In *Document Understanding Conference (DUC)*, New Orleans, LA, USA.

RADEV D. R., JING H., STYŚ M. & TAM D. (2004). Centroid-based summarization of multiple documents. *Information Processing and Management*, **40**(6), 919–938.

SPÄRCK JONES K. & GALLIERS J. R. (1996). *Evaluating Natural Language Processing Systems : An Analysis and Review*. Springer.

TEUFEL S. & MOENS M. (1997). Sentence extraction as a classification task. In *ACL/EACL Workshop on Intelligent Scalable Text Summarization*, p. 58–65, Madrid, Spain.

## Annexe

- S0 La police a procédé aux arrestations lors d'une manifestation à Lhassa, coïncidant avec le 49e anniversaire du départ forcé du dalaï lama.
- S1 Une soixantaine de moines bouddhistes ont été arrêtés lundi à Lhassa, la capitale du Tibet, à l'occasion d'une manifestation coïncidant avec le 49e anniversaire du départ forcé du dalaï lama, a affirmé mardi Radio Free Asia (RFA).
- S2 Entre 50 et 60 manifestants ont été arrêtés par la police par les forces de l'ordre, qui ont également bloqué les routes et encerclé les monastères pour empêcher les manifestations de se propager.
- S3 Cependant, onze personnes ont réussi à protester dans le centre de Lhassa avant d'être arrêtées, selon les mêmes sources citées par RFA.
- S4 Des responsables de la police et des affaires religieuses à Lhassa ont refusé de s'exprimer.
- S5 Le dignitaire religieux de 72 ans, qui a fui le Tibet en 1959 après l'échec d'un soulèvement anti-chinois, a abandonné ses revendications d'indépendance, se bornant à réclamer "une large autonomie" pour sauvegarder la langue, la culture et l'environnement de ce territoire himalayen.
- S6 La Chine, qui en a pris le contrôle à partir de 1950 -avant d'y mener une sanglante répression- n'a cessé de rejeter ces demandes qualifiées par le dalaï lama de diplomatie de la "voie moyenne".
- S7 Depuis lundi, des moines bouddhistes manifestent au Tibet et dans les régions avoisinantes, à l'occasion du 49e anniversaire du soulèvement de Lhassa, qui a conduit à l'exil du dalaï-lama.
- S8 Depuis Dharamsala, dans le nord de l'Inde, le dalaï-lama a demandé à Pékin de "renoncer à l'usage de la force" contre les manifestants.
- S9 Son porte-parole a jugé sans fondement les accusations chinoises selon lesquelles il aurait fomenté les manifestations violentes.
- S10 Ce nouvel embrasement, dans une région sensible, sous contrôle chinois depuis 1951, devrait accentuer la pression que subit déjà le gouvernement chinois pour améliorer les droits de l'homme, comme il s'est engagé à le faire en obtenant l'organisation des JO de Pékin, dont l'ouverture aura lieu dans cinq mois.
- S11 Le gouvernement chinois a proposé d'indemniser les familles des civils qui ont, selon lui, été tués lors des violences dans la capitale tibétaine ce mois-ci, a rapporté vendredi soir l'agence de presse officielle chinoise Chine nouvelle.
- S12 Selon le décompte du gouvernement chinois, 18 civils ont été tués le 14 mars lors de manifestations contre la tutelle chinoise à Lhassa, au cours desquelles des manifestants ont lancé des pierres en direction des forces de l'ordre, brûlé et pillé des magasins.
- S13 Les familles des victimes recevront 200.000 yuans ( 18.000 euros), a indiqué Chine nouvelle sur la foi d'une circulaire du gouvernement régional du Tibet.
- S14 "Des mesures sont prises pour aider les gens à réparer leurs maisons et leurs magasins détruits pendant les troubles ou à en construire d'autres", précise la circulaire selon Chine nouvelle.
- S15 Toute personne blessée lors des émeutes pourra être soignée gratuitement, a ajouté Chine nouvelle.
- S16 Le bilan officiel de deux semaines de violences au Tibet et dans l'ouest de la Chine est de 19 morts, mais le gouvernement tibétain en exil fait état de 140 morts.
- S17 La répression des émeutes par les autorités chinoises a provoqué des protestations internationales à l'approche des Jeux olympiques de Pékin.

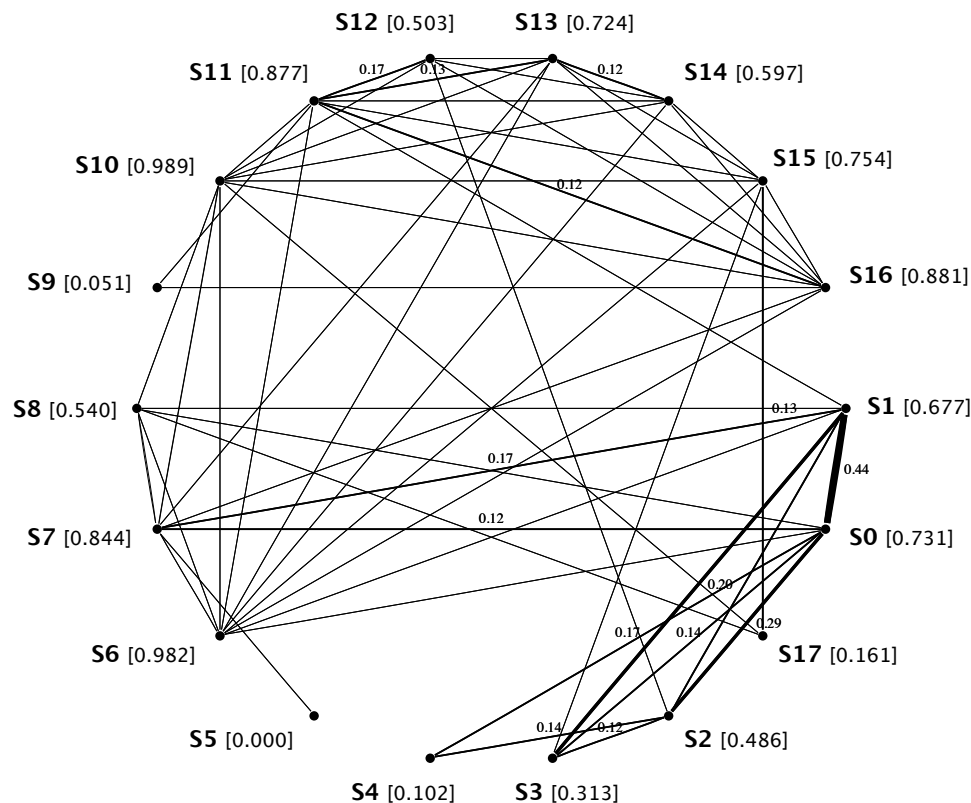


TABLE 2 – Exemple de graphe construit à partir d'un cluster d'articles journalistiques traitant de l'actualité au Tibet. Les scores TextRank de chaque segment sont montrés entre crochets.