Le résumé par classification

Principes et applications

Aurélien Bossard [1], **Émilie Guimier De Neef** [2]

1. *Orange Labs*
Avenue Pierre Marzin
22300 Lannion
aurelien.bossard@gmail.com

2. *emilie.guimierdeneef@orange-ftgroup.com*

RÉSUMÉ. *Cet article présente en détail un système de résumé automatique fondé sur la méthode CBSEAS, qui utilise la classification de phrases en classes informationnelles afin de générer des résumés. Cette méthode a déjà été décrite, mais pour la première fois, les résumés sont fondés sur une classification des phrases entièrement non supervisée. L'article présente également une méthode de résumé de nouveauté, et son évaluation sur un corpus libre français. Enfin, nous proposons une expérience réalisée sur un corpus composé d'extraits de Wikipédia, afin d'évaluer la performance d'une méthode purement statistique sur des données autres que les actualités, sur lesquelles se concentrent traditionnellement les systèmes de résumé automatique.*

ABSTRACT. *This article introduces an automatic summarization system based on CBSEAS, a method that uses sentence clustering in informational classes in order to produce automatic summaries. Other articles already dealed with CBSEAS, but for the first time, the summaries are based on an entirely supervized sentence classication algorithm. The article also introduces an update summarization method and its evaluation on an open French corpus. The paper presents an experiment conducted on Wikipédia aggregates in order to evaluate CBSEAS on a different task than newswire summarization.*

MOTS-CLÉS : *résumé automatique, résumé de mise à jour, classification de phrases, extraction de phrases.*

KEYWORDS: *automatic summarization, update summarization, sentence clustering, sentence extraction.*

DOI:10.3166/DN.15.2.11-39

1. Introduction

Les approches de résumé automatique multidocuments utilisant une classification des phrases en fonction de leur vocabulaire se sont multipliées ces dernières années. Celles-ci sont en effet particulièrement efficaces pour éliminer la redondance dans les résumés générés, mais également pour sélectionner des fragments de texte pertinents. Des campagnes d'évaluation récentes ont montré qu'elles sont performantes sur des tâches de résumé d'actualité ou de résumé d'opinions issues de blogs. Nous étudions dans cet article leur adéquation à des types de documents et des types de résumés différents, afin d'en évaluer la généricité.

La première partie de cet article est consacrée à l'état de l'art des méthodes de résumé atomatique. La seconde partie présente un système de résumé automatique multidocument par classification, CBSEAS. Nous détaillons ensuite l'adaptation de CBSEAS à la révision de résumé – ou résumé de mise à jour. Nous présentons ensuite divers cas d'étude sur des types de contenu différents, qui nous permettent d'évaluer la généricité du système CBSEAS, avant d'aborder nos conclusions sur les expériences réalisées et une discussion sur la généricité des systèmes de résumé automatique.

2. État de l'art

Le résumé automatique est étudié depuis le début du traitement des données textuelles. Très vite, les méthodes génératives ont montré leurs limites. Ces approches fortement dépendantes de la langue nécessitent en effet des ressources linguistiques complexes. Récemment, les recherches de Marcu (1998) ont tenté d'analyser la structure rhétorique pour sélectionner des phrases pertinentes, mais cette méthode est toujours limitée à des domaines applicatifs spécifiques.

Depuis les années 1950 (Luhn, 1958), la recherche en résumé automatique s'est concentrée sur l'extraction de phrases importantes – la création d'extracts – plutôt que sur la génération d'abstracts. Les phrases extraites doivent constituer un texte cohérent, fidèle aux idées/informations exprimées dans les documents d'origine. L'extraction de phrases est généralement réalisée en calculant un score pour chaque phrase, et en extrayant les mieux classées afin de produire un résumé. Le nombre de phrases ou de mots dans le résumé peut être déterminé à l'avance, mais peut également être calculé dynamiquement en utilisant un pourcentage de compression – par exemple 10 % des documents d'origine.

(Edmundson, Wyllys, 1961) a défini des indices de surface qui peuvent être utilisés afin de déterminer l'importance d'une phrase. Ces indices comprennent la position d'une phrase, la présence de mots-clés et de mots du titre, ainsi que d'entités nommées. Ils sont encore utilisés de nos jours dans la majorité des systèmes de résumé automatique, comme celui de (Kupiec *et al.*, 1995), qui les combine à un algorithme d'apprentissage. Cependant, ils sont limités car ils ne prennent pas en compte le contenu global du document.

D'autres systèmes se fondent sur la fréquence des mots en corpus. C'est le cas de (Luhn, 1958), mais également de (Radev *et al.*, 2002), qui a introduit le *tf.idf* comme indice de pouvoir informatif des unités lexicales dans sa méthode du centroïde. Les phrases extraites sont celles qui possèdent le plus de mots fortement informatifs.

Ces méthodes sont efficaces pour générer des résumés dits « centraux », c'est-à-dire qui reflètent le contenu global des documents. En revanche, elles ne garantissent pas la diversité informationnelle des résumés, ou le fait que chacune des phrases d'un résumé traite d'une information différente. Cette diversité informationnelle est un critère important de sélection des phrases : un résumé doit contenir toutes les informations importantes.

L'algorithme MMR –*Maximal Margin Relevance*– (Carbonell, Goldstein, 1998) résoud le compromis entre diversité et centralité grâce à la fonction de score présentée en équation (1). Un score de diversité est soustrait au score de centralité de chaque phrase. À chaque itération de l'algorithme, la phrase la mieux classée est extraite dans le résumé.

$$MMR = \underset{P_i \in D \setminus S}{\operatorname{argmax}} \left[\lambda centr(P_i) - (1-\lambda) \underset{P_j \in S}{\operatorname{argmax}}\ sim_2(P_i, P_j) \right] \qquad (1)$$

où $centr$ est un score de centralité, D l'ensemble des phrases à résumer, S les phrases déjà sélectionnées, et λ le facteur de nouveauté.

(Erkan, Radev, 2004) s'est appuyé sur les avancées récentes dans le domaine des réseaux sociaux afin d'utiliser les similarités entre phrases comme critère de centralité. Ainsi, ce ne sont plus des mots isolés qui mesurent la centralité d'une phrase, mais les similarités entre phrases, qui prennent en compte les cooccurrences. La centralité est donc plus fonction des concepts évoqués dans une phrase que d'unités lexicales isolées. La méthode LexRank tirée de ces travaux et fondée sur l'analyse du graphe des phrases à résumer, est présentée en section 3.3.1.

D'autres travaux utilisent la hLDA – *hierarchical Latent Dirichlet Allocation* – afin de générer des résumés (Darling, Song, 2011). Les phrases à résumer sont vues comme des chemins dans un arbre hiérarchique de thèmes. Le processus de résumé revient alors à sélectionner les phrases qui maximisent la couverture des thèmes. Ce processus peut tout de même générer des résumés redondants, et les auteurs emploient une méthode d'élimination de la redondance fondée sur l'utilisation de paramètres fixés empiriquement.

Au-delà de l'introduction d'un nouveau type de centralité – la centralité locale –, qui vise à évaluer la centralité d'une phrase en rapport avec son thème, nous défendons dans cet article une méthode de résumé automatique – CBSEAS – qui élimine la redondance de manière entièrement non supervisée.

3. CBSEAS : Clustering-Based Sentence Extractor for Automatic Summarization

Un corpus multidocument est composé d'une multitude d'informations qui constituent la diversité informationnelle des documents à résumer. Chaque information est véhiculée par une ou plusieurs phrases. Ce phénomène de redondance est d'autant plus prononcé que les documents composant le corpus à résumer traitent de thèmes proches, voire d'un seul et même thème. Identifier automatiquement les phrases ayant trait à un même thème permet d'obtenir des classes thématiques, qui apportent des informations supplémentaires et donc un gain de précision pour l'extraction de phrases.

La méthode CBSEAS, déjà décrite dans (Bossard *et al.*, 2010), vise justement à regrouper les phrases en classes thématiques avant la phase d'extraction de phrases. Nous avons implémenté cette méthode dans un système de résumé automatique qui utilise des techniques de regroupement et d'extraction de phrases sensiblement différentes de celles présentées dans (Bossard *et al.*, 2010). Nous avons en effet ajouté un indice de qualité de clustering, qui permet de déterminer automatiquement le nombre de classes thématiques optimales et rend donc l'approche de classification entièrement non supervisée, ainsi que de nouveaux paramètres (similarité au titre, importance des classes thématiques) dans le calcul du score d'une phrase. Cette section décrit ce nouveau système de résumé, tandis que les sections suivantes sont consacrées à son évaluation sur différents cas d'étude.

3.1. La méthode CBSEAS : principes généraux

La méthode CBSEAS est dédiée au résumé automatique multidocument. Elle vise à identifier et à regrouper les phrases qui traitent d'une même composante informationnelle, afin d'apporter trois nouveaux éléments au calcul de résumés :

– premièrement, l'évaluation de l'importance d'une composante informationnelle en fonction du nombre de phrases qui la composent. CBSEAS est fondée sur l'hypothèse qu'au sein d'un corpus multidocument, la redondance est un bon indice de pertinence ;

– deuxièmement, la pertinence d'une phrase vis-à-vis de la composante informationnelle qu'elle représente (et non plus seulement vis-à-vis du contenu global des documents à résumer) ;

– enfin, l'élimination de la redondance en ne sélectionnant qu'une phrase par composante informationnelle.

Dans les sections qui suivent, nous nous attachons à montrer que chacun de ces éléments apporte une précision supplémentaire à la sélection des phrases pour le résumé, au travers d'expériences réalisées sur un corpus français d'évaluation de résumés d'actualité.

3.2. Implémentation du regroupement de phrases

Parmi les méthodes de classification non supervisée, une des plus utilisées est la méthode des k-moyennes (*k-means*) (MacQueen, 1967) et ses variantes. Celle-ci présente un avantage certain par rapport à d'autres méthodes, comme les modèles de mélanges gaussiens : la classification est fondée sur une matrice de similarités entre les éléments à classifier. Notre problème – la classification de phrases – nécessite une modélisation linguistique fine. En effet, contrairement à la classification de documents, une tâche plus classique du TAL, les objets sont classifiés selon des caractéristiques en nombre limité.

Les classifieurs automatiques fondés sur la méthode des k-moyennes nous permettent d'expérimenter différentes mesures de similarité, que ce soit au niveau des phrases (Bossard, Guimier De Neef, 2011a) ou des unités lexicales (Bossard, 2010). Cependant, dans cet article, nous ne présentons qu'une seule mesure de similarité, fondée sur une approche « sac de mots », qui s'est jusqu'ici révélée la plus performante pour la classification de phrases (*cf* Tableau 1, où les scores de résumé obtenus par un regroupement de phrases selon différentes mesures de similarité et d'édition sont comparés à une mesure sac de mots pondérée par le tf.idf)

Tableau 1. Scores ROUGE de CBSEAS selon différentes mesures de similarité et d'édition pour réaliser le regroupement de phrases : cosinus unigramme, bigramme, bigramme à trou, Levenshtein, Jaro-Winkler, Smith-Waterman

	cos-1	cos-2	cos-2-sk4	Leven.	J-W	Sm.-Wat
ROUGE1	0.39304	0.37843	0.36153	0.37950	0.37625	0.38257
ROUGE2	0.12815	0.11871	0.10030	0.12338	0.11922	0.12503
ROUGE-SU4	0.14741	0.14056	0.12771	0.14180	0.13904	0.14390

D'autres classifieurs sont fondés sur une matrice de similarité ; nous avons cependant choisi une variante de l'algorithme des k-moyennes, pour des raisons que nous détaillons en section 3.2.2.

3.2.1. Similarités entre phrases

Dans un premier temps, le corpus à résumer est lemmatisé. Nous utilisons pour cela TiLT-Lemmatiseur (Heinecke *et al.*, 2008), qui assure la désambiguïsation morpho-syntaxique des segments de la phrase au moyen d'une grammaire hors contexte de chunking. Pour le français, la grammaire comporte autour de 2 000 règles et s'appuie sur un lexique du français de plus de 150 000 lemmes enrichi de près de 350 000 entités nommées, soit 500 000 entrées. TiLT-Lemmatiseur produit des étiquettes conformes à celles proposées pour le français dans TreeTagger [1] et avait été évalué dans le cadre de la campagne GRACE (Adda *et al.*, 1999) avec une précision supérieure à 95 %.

1. http://www.ims.uni-stuttgart.de/~schmid/french-tagset.html

Un poids est alors calculé pour chacun des couples (lemme, catégorie morpho-syntaxique) identifiés. Nous différencions donc « OR, *nom commun* » de « OR, *conjonction de coordination* ». Les poids calculés dépendent en grande partie du type de corpus en entrée (monodocument ou multidocument), et de l'application visée (résumé standard, résumé de mise à jour...). Nous présentons dans les sections relatives à chacune de nos expériences les calculs associés au poids des couples (lemme, catégorie).

Les pronoms, prépositions, déterminants et conjonctions de coordination sont éliminés, du fait de leur faible pouvoir informatif et discriminant. Les mots appartenant aux autres catégories morpho-syntaxiques (noms, adverbes, adjectifs, verbes, nombres) sont utilisés afin de calculer les similarités entre phrases, selon une mesure *cosinus* pondérée par le poids des couples (lemme, catégorie), présentée en équation 2. Nous avons précédemment montré que cette mesure est celle qui nous a permis d'obtenir les meilleurs résumés en sortie (Bossard, Guimier De Neef, 2011a).

$$cosinus_{ponderee}(p1, p2) = \frac{\sum_{l \in p1, p2} poids(l)^2}{\sqrt{\sum_{l_i \in p1} poids(l_i)^2} \times \sqrt{\sum_{l_j \in p2} poids(l_j)^2}} \quad (2)$$

où p1 et p2 sont des ensembles de couples (lemme, catégorie morpho-syntaxique).

3.2.2. Algorithme de classification : fast global k-means

Nous avons choisi d'utiliser un classifieur fondé sur la méthode des k-moyennes, *fast global k-means* (López-Escobar *et al.*, 2006). *Fast global k-means* est une variante incrémentale des k-moyennes, qui vise à résoudre le problème du choix des k centres de classe initiaux posé par l'algorithme d'origine. L'incrémentalité de *fast global k-means* est également intéressante dans le but de générer des résumés de mise à jour, une tâche par définition incrémentale.

$$Ind_{D-B} = \frac{1}{N} \sum_{i=1}^{N} \frac{E(C_i) - E(C_j)}{S(C_i, C_j)} \quad (3)$$

où :

- N est le nombre de clusters ;
- $E(C_i)$ est une mesure d'éparpillement du cluster C_i (*cf* [4]) ;
- $S(C_i, C_j)$ est une mesure de séparation des clusters C_i et C_j (*cf* [5]) .

$$E(C_i) = \frac{1}{N_i} \sum_{j=1}^{N_i} ||x_j - G_i|| \quad (4)$$

$$S(C_i, C_j) = ||G_i - G_j|| \quad (5)$$

– G_i est le barycentre de C_i ;
– N_i le nombre d'éléments de C_i ;
– x_j les éléments d'un cluster.

Fast global k-means crée d'abord une classe qui contient tous les éléments à classifier. À chaque itération, l'algorithme ajoute une nouvelle classe avec pour seul élément celui qui, parmi tous les éléments, est le moins représentatif de sa classe – donc le plus éloigné du barycentre de celle-ci. Chaque élément est alors placé dans la classe dont il est le plus proche du barycentre, et le centre de chaque classe est recalculé. L'algorithme s'arrête lorsque le nombre de classes (k) demandé par l'utilisateur est atteint.

3.2.3. Trouver le meilleur k

Le principal inconvénient de cet algorithme réside dans le fait que le k doit être connu à l'avance, contrairement à d'autres classifieurs, fondés par exemple sur des modèles de mélanges de gaussiennes. Il est cependant possible d'estimer le k pour lequel la classification est optimale. Il faut pour cela avoir recours à un indice de qualité de *clustering*, qui permette d'évaluer et comparer la qualité de partitionnements suivant différents nombres de classes. Différents indices de ce type existent, dont les plus connus sont les indices de Dunn (1973), Davies and Bouldin (1979) et Silhouette (Rousseeuw, 1987). De ces trois indices, Davies-Bouldin est celui qui possède la complexité la plus faible. Étant donné que ces indices sont calculés pour chacun des k possibles, une complexité élevée est prohibitive.

De plus, l'indice de Davies-Bouldin favorise les groupes hypersphériques, et est donc particulièrement bien adapté pour une utilisation avec la méthode des k-moyennes (Guérif, 2006). Cet indice est présenté en équation 3.

L'indice est négativement corrélé à la qualité de partitionnement. Un indice faible suppose donc une qualité de partitionnement élevée. Nous avons constaté lors d'expériences préliminaires qu'au-delà d'un certain nombre de classes, variable suivant les données, l'indice atteignait un seuil puis décroissait continuellement jusqu'à atteindre sa plus basse valeur pour un nombre de classes égal au nombre d'éléments. Ceci est illustré par la Figure 1.

L'observation des distances entre phrases a révélé que leur répartition n'était pas adaptée aux différents indices de qualité de partitionnement. On voit nettement, sur la Figure 2, que trop peu de distances ont une valeur moyenne, comparées aux distances élevées. La prise en compte des seules unités lexicales porteuses de sens est responsable de ce phénomène. En effet, les variables représentant les phrases sont trop peu nombreuses. Une solution pourrait consister à plafonner le nombre de classes à la moitié des phrases. Cependant, dans le cas où les données sont réellement éparpillées, cette solution apparaît peu robuste.

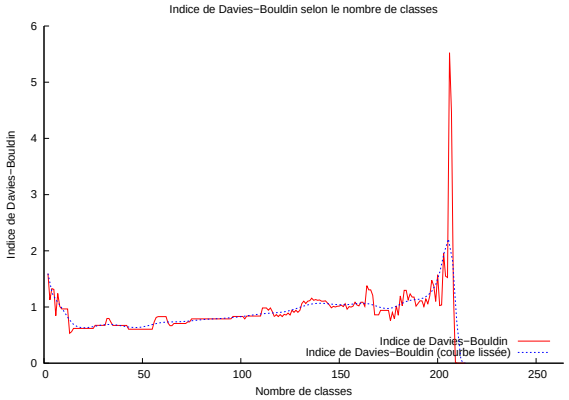

Figure 1. Indice de Davies-Bouldin suivant le nombre de classes

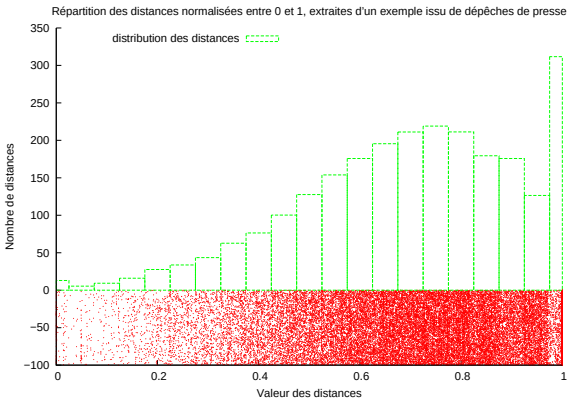

Figure 2. Répartition des distances entre phrases sur un exemple issu d'un corpus d'actualités

Nous avons résolu ce problème en optant pour une distance entre phrases tenant compte de toutes les unités lexicales, tout en supprimant leur pondération par le *tf.idf*. Il en résulte des indices de qualité de partitionnement plus conformes à nos attentes (*cf*

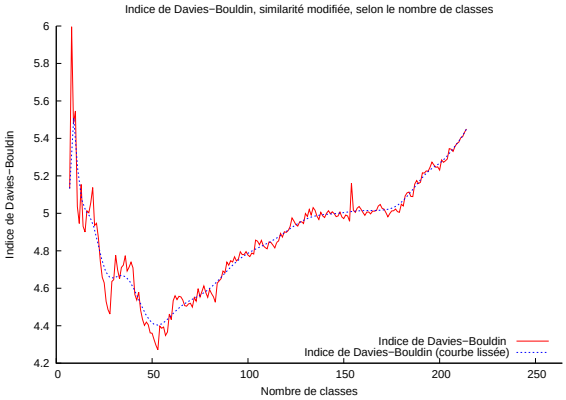

Figure 3. Indice de Davies-Bouldin après correction des distances entre phrases

Figure 3) : l'indice, passé un certain nombre de classes, et donc un trop grand nombre d'éclatement de classes, ne fait plus que croître jusqu'à ce qu'il ne soit plus défini – il suffit pour cela que deux classes ne contiennent plus qu'une même phrase.

3.3. Evaluation de la pertinence des phrases

Après avoir regroupé les phrases, notre système extrait les phrases jugées les plus pertinentes. La pertinence d'une phrase est fonction de quatre scores : sa centralité vis-à-vis de sa composante informationnelle et la pertinence de cette dernière, sa centralité vis-à-vis du contenu global des documents, et sa similarité au titre de son document, dans le cas où les documents sont titrés. Dans cette section, nous justifions ces scores et détaillons leur implémentation.

3.3.1. Centralité globale

La centralité, dans le domaine du résumé automatique, correspond à la proximité d'un segment textuel aux informations les plus importantes des documents d'origine. Tous les systèmes de résumé automatique par extraction implémentent une mesure de centralité globale afin de décider des segments textuels à extraire. Ces mesures peuvent être fonction d'une comparaison du lexique de la phrase visée avec le vocabulaire global des documents, de la similarité aux autres phrases, ou encore de la position de la phrase dans l'articulation d'un discours. Les mesures de centralité sont présentées plus en détail en §2.

Nous avons choisi d'implémenter une approche à base de graphes. En effet, ces approches sont les plus précises parmi les approches fondées sur le vocabulaire. Prendre en compte uniquement le vocabulaire permet de s'affranchir de toute hypothèse sur l'articulation logique du discours. L'approche est donc généralisable à tous documents dans lesquels la fréquence d'un fragment textuel témoigne de son importance.

Nous avons implémenté la mesure de centralité LexRank, tirée de (Erkan, Radev, 2004). Elle vise à faire ressortir les phrases qui ont le plus de « prestige » [2], c'est-à-dire celles qui sont fortement liées à des phrases elles-mêmes dotées d'un fort prestige. Le calcul est itératif : le prestige d'une phrase dépend en effet de celui des phrases qui l'entourent. A chaque itération, le score de chaque phrase est calculé selon la formule décrite en équation 6. Le prestige de chaque phrase est calculé jusqu'à convergence, soit jusqu'à ce qu'aucun score ne varie plus qu'un δ fixé à $1e^{-9}$.

$$prest(u) = \frac{d}{N} + (1-d) \sum_{v \in adj[u]} \frac{sim(u,v) \times prest(v)}{\sum_{z \in adj[v]} sim(z,v)} \quad (6)$$

où :

- N est le nombre total de nœuds,
- $adj[u]$ est l'ensemble des nœuds adjacents à u.

3.3.2. Centralité locale

L'apport majeur de notre système de résumé automatique est l'évaluation de la centralité d'une phrase par rapport à sa composante informationnelle. Cette nouvelle mesure s'ajoute à la centralité globale, et permet ainsi d'apporter une vue différente de la centralité avec une granularité plus fine qu'un score fondé sur l'ensemble des phrases d'origine.

La centralité locale est calculée par le même algorithme que la centralité globale. La différence réside dans la construction du graphe : seules les phrases classées dans une même composante informationnelle sont utilisées. De même, le poids des unités lexicales est calculé différemment : si pour la centralité globale le poids d'une unité lexicale est fonction de sa fréquence dans le corpus [3], pour la centralité locale le poids d'une unité lexicale est un tf.idf où la référence est la classe informationnelle. Cela permet de prendre en compte l'importance relative des termes dans une composante informationnelle donnée.

Cette mesure pourrait toutefois provoquer un biais : sélectionner des phrases centrales localement, mais hors sujet globalement. Pour éviter cet écueil, la mesure de centralité locale est pondérée par une mesure d'importance des classes informationnelles, décrite dans la section suivante.

2. Prestige est entendu ici dans le sens qu'on lui donne dans le domaine des réseaux sociaux.

3. Dans certains cas, on utilisera un tf.idf, dans d'autres simplement la fréquence des termes ; les poids liés aux unités lexicales sont détaillés dans les sections relatives aux différents cas d'étude.

3.3.3. Pertinence globale d'une composante informationnelle

Nous faisons l'hypothèse que dans un corpus multidocument, la redondance d'une information est un bon indice de sa pertinence. Par conséquent, nous considérons que plus une classe informationnelle est peuplée, plus elle est importante dans le corpus à résumer. Le score de pertinence d'une classe informationnelle sert de facteur multiplicateur au score de pertinence locale. Le score d'une classe informationnelle C_i est donc calculé comme décrit en équation 7 :

$$pert(C_i) = \frac{|C_i|}{\max\limits_{C_j \in C} |C_j|} \tag{7}$$

3.3.4. Similarité au titre

Parfois, les termes utilisés dans le titre d'un document peuvent être de bons indicateurs de pertinence : les phrases qui contiennent ces termes ont une probabilité plus importante d'être centrales. Les études menées par (Edmundson, Wyllys, 1961) sur les indicateurs de pertinence vont dans ce sens, et le titre est encore utilisé en tant qu'indicateur de pertinence dans la plupart des systèmes de résumé modernes. Cependant, sur certains types de documents, le titre peut être soit trop abstrait et donc faiblement indicatif du contenu, soit trop spécialisé et éluder des parties importantes du discours. Nous avons donc choisi de rendre optionnelle cette mesure de similarité au titre. Celle-ci se calcule de la même manière que les similarités entre phrases, présentées en section 3.2.1.

3.3.5. Longueur des phrases

Les diverses expériences menées durant les campagnes d'évaluation de résumés TAC 2008 et TAC 2009 ont montré que la forme d'un résumé influait autant que le fond sur sa perception par un utilisateur. Il ne suffit donc pas de générer des résumés informatifs ; il faut également que ceux-ci soient lisibles et leur discours bien construit. Si la construction du discours est une problématique compliquée qui se résoud notamment par des modules de réordonnancement des phrases, la lisibilité peut être améliorée en sélectionnant des phrases dont l'aspect est plus propice à une lecture rapide. La longueur des phrases est un des facteurs connus de complexité de lecture, et constitue donc un bon indice afin d'évaluer la lisibilité (Flesch, 1948).

Il existe des indices fondés sur la complexité du vocabulaire (mesurée par le nombre de syllabes des mots) que nous n'avons pas retenus en raison du risque de conflit avec les mesures de centralité que nous utilisons : un mot pourrait être à la fois central et complexe. D'autres indices existent, fondés notamment sur la complexité syntaxique apparente (*e.g.* nombre et type des subordonnants). Ces indices rendraient cependant notre système, que nous voulons le plus générique possible, fortement dépendant de la langue.

Utiliser le $\log$ de la longueur des phrases comme score de complexité permet de ne pas avantager des phrases légèrement plus petites, mais moins informatives que

d'autres. Le score de complexité que nous utilisons : $\log(N)$, où N est le nombre de mots d'une phrase, permet un bon compromis entre informativité et lisibilité.

3.3.6. Addition des scores

Une fois les différents scores calculés, le système affecte un score à chaque phrase : une somme pondérée des scores décrits dans les sections 3.3.1 à 3.3.4. Le score d'une phrase p appartenant à une classe informationnelle C est calculé d'après la formule donnée en équation 8 :

$$score_{global}(p) = \alpha centr_g(p) + \beta pert(C) centr_l(c) + \gamma score_t(p) \qquad (8)$$

où :

- $centr_g$ est la centralité globale ;
- $centr_l$ la centralité locale ;
- $pert$ la pertinence d'une composante informationnelle ;
- $score_t$ la similarité au titre.

Si l'utilisation de phrases courtes pour constituer le résumé est préférée, le score est divisé par le score de longueur, présenté en section 3.3.5.

Il est possible de paramétrer finement ces poids en couplant un système de résumé à un algorithme d'optimisation. Nous avons montré dans (Bossard, Rodrigues, 2011) qu'en disposant d'un corpus d'entraînement [4], un algorithme génétique permet une augmentation des scores automatiques ROUGE (C.-Y. Lin, 2004) de près de 10 % sur les données de la campagne d'évaluation TAC 2008 [5]. Cependant, de tels algorithmes sont gourmands en ressources, et les poids ainsi calculés ne sont valables que pour des types de résumé et de document particuliers. C'est pourquoi dans les expériences décrites dans cet article, nous avons paramétré empiriquement et manuellement les différents poids relatifs aux scores des phrases.

3.4. Sélection des phrases

Lorsqu'un score a été attribué à chacune des phrases, le système de résumé sélectionne les phrases de manière à générer le meilleur résumé possible. Il s'agit de sélectionner les meilleures phrases possibles selon les critères de centralité, tout en satisfaisant les contraintes de redondance – ne sélectionner qu'une phrase par classe informationnelle – et de taille. La taille peut être fixée de deux manières différentes : soit en nombre de mots *a priori* – *e.g.* un maximum de 150 mots – soit en pourcentage

4. Il faut donc disposer de plusieurs groupes de documents, de résumés manuels ainsi que d'un système d'évaluation automatique de résumés.

5. Les campagnes TAC, pour *Text Analysis Conference*, sont organisées annuellement par le *National Institute of Science and Technology.* : http://www.nist.gov/tac

de compression des textes d'origine – la contrainte est alors rapportée à un nombre de mots maximum.

L'algorithme de sélection de phrases agit comme suit : à chaque itération, il examine toutes les classes informationnelles marquées « S » (les classes sont toutes marquées « S » par défaut). Il sélectionne parmi ces classes la phrase dont le score est maximal et dont la taille respecte la contrainte de taille du résumé. La classe dont est issue cette phrase est marquée « NS ». Toutes les classes dont aucune phrase ne respecte la contrainte de taille sont marquées « NS ». L'algorithme est répété jusqu'à ce que toutes les classes soient marquées « NS ».

3.5. Ordonnancement des phrases

Les phrases sélectionnées doivent alors être ordonnées afin d'être présentées à l'utilisateur sous la forme d'un résumé le plus cohérent possible. De nombreux travaux ont porté sur l'ordonnancement de phrases, et les méthodes vont de l'analyse temporelle (Z. Lin *et al.*, 2008) ou rhétorique (Rutledge *et al.*, 2000) à des techniques plus frustres tentant de maximiser la continuité dans le vocabulaire des phrases (Wang, 2002), utilisant la spécificité de la modélisation du corpus en classes informationnelles et la position des phrases dans leur document – *cf* Barzilay *et al.* (2002) et Bossard and Guimier De Neef (2011b) – ou encore utilisant simplement la date des documents et la position des phrases (Saggion, Gaizauskas, 2004). Ces dernières techniques présentent l'avantage d'être robustes et portables.

Nous avons montré dans (Bossard, Guimier De Neef, 2011b) que pour le résumé d'actualité, ordonner les phrases selon leur score était une meilleure stratégie que d'utiliser leur position et la modélisation en classes informationnelles. Les résumés ordonnés selon cette stratégie satisfont plus les lecteurs. Notre système étant principalement dédié au résumé d'actualités, nous avons voulu privilégier un ordonnancement des phrases soit temporel, soit selon leur caractère marquant. Nous avons donc éliminé la dernière solution robuste, qui consiste à ordonner les phrases de manière à maximiser le vocabulaire commun entre phrases juxtaposées, et privilégié l'ordonnancement selon le score calculé par le système.

4. Révision de résumés (ou mise à jour de résumés)

La mise à jour de résumé apparaît comme une discipline nouvelle dans le domaine du résumé automatique. Introduite pour la première fois dans une campagne d'évaluation internationale en 2007 par la campagne DUC [6], la mise à jour de résumé vise à comparer deux corpus portant sur des intervalles de temps différents – nommément « corpus initial » et « corpus de mise à jour » –, et à générer un résumé des informations nouvelles du corpus le plus récent.

6. Les campagnes DUC, pour *Document Understanding Conference.*, aujourd'hui renommées en TAC, ont été organisées par le *National Institute of Science and Technology* jusqu'en 2007 : http://duc.nist.gov

Au-delà de la sélection de phrases centrales, la problématique principale posée par le résumé de mise à jour réside dans la distinction entre phrases porteuses d'informations nouvelles et phrases porteuses d'informations anciennes. Cette problématique nouvelle complexifie de manière importante la tâche de résumé automatique standard, comme l'ont montré les campagnes d'évaluation DUC 2007, TAC 2008 et TAC 2009 (Dang, Owczarzak, 2008).

Le problème de la détection de nouveauté peut être abordé de différentes manières : certains auteurs (Galanis, Malakasiotis, 2008) utilisent un seuil empirique entre phrases du corpus de mise à jour et phrases du corpus initial. Toute phrase ayant une similarité avec une phrase du corpus initial au-dessus de ce seuil est considérée comme non porteuse de nouveauté.

D'autres auteurs suppriment les phrases qui maximisent la similarité au jeu de documents initial jusqu'à ce que la similarité globale entre ce dernier et le document de mise à jour descende en dessous d'un seuil prédéfini (He *et al.*, 2008). Ne restent alors dans le jeu de documents de mise à jour que les phrases les plus différentes du corpus d'origine.

Boudin *et al.* (2008) présentent une variante de la méthode MMR – décrite en §2 – dédiée au résumé de mise à jour. Les phrases des documents d'origine sont ajoutées au jeu des phrases déjà sélectionnées dans le résumé, et le poids de la similarité à ce jeu de phrases dans le score MMR est augmenté afin de favoriser les phrases porteuses de nouveauté.

Une autre méthode, introduite dans (Varma *et al.*, 2009), vise à évaluer la nouveauté d'un mot. Le facteur de nouveauté (fn) d'un mot dans un document publié à une date t dépend de son nombre d'occurrences dans les documents antérieurs et postérieurs. Sa formule est présentée en équation 9. Le facteur de nouveauté des mots est utilisé pour évaluer la nouveauté d'une phrase. Cette méthode a prouvé son efficacité sur les évaluations de TAC 2009.

$$fn(w) = \frac{|nd_t|}{|pd_t| + |D|} \tag{9}$$

$nd_t = d : w \in d \wedge t_d\, t$
$pd_t = d : w \in d \wedge t_d \leq t$
$D = d : t_d\, t$

Contrairement aux premières approches présentées, nous voulons nous affranchir de seuils ou poids empiriques pour la détection de la nouveauté. De plus, nous voulons une approche généralisable au résumé différentiel, qui vise non pas à résumer la nouveauté, mais les informations différentes entre deux corpus. Notre méthode doit donc être indépendante de la date des documents.

4.1. Méthode de détection de la nouveauté

Notre système de résumé standard, présenté dans la section précédente, regroupe automatiquement les phrases les plus similaires. En d'autres termes, il crée différentes classes pour des phrases distantes sémantiquement. Notre méthode de classification peut aussi être utilisée pour classer les phrases en deux groupes :

– celles qui portent des informations nouvelles ;

– celles qui portent des informations déjà connues.

Notre méthode est donc fondée sur l'hypothèse selon laquelle les phrases des documents de mise à jour porteuses d'informations initiales partagent le même vocabulaire que les phrases des documents initiaux. Elle s'affranchit donc de l'utilisation d'indices discursifs tels que finalement, qui permettrait de confirmer que la seconde phrase de la Figure 4 est porteuse d'une mise à jour de l'information portée par la première phrase. L'utilisation de tels indices permettrait une précision accrue, mais nécessiterait un apprentissage spécifique à une tâche (ici le résumé de mise à jour), à un langage et à un type de documents uniques.

« Nicolas Hulot fait planer la menace d'une candidature dissidente. »
« Après avoir brandi la menace d'une candidature dissidente, Nicolas Hulot a **finalement** promis de rester dans le giron d'EELV. »

Source : L'Express

Figure 4. Illustration de l'importance des indices discursifs dans la détection de mise à jour

Avant d'identifier les phrases porteuses de nouveauté, il est nécessaire de modéliser les informations du corpus initial. Il sera alors possible de confronter les informations des documents de mise à jour avec celles des documents initiaux. La première étape de notre algorithme de détection de mise à jour consiste donc à classifier les phrases des documents initiaux (D_I) en k_I classes, selon la méthode présentée en section 3.2.

Le modèle initial (M_i) est le graphe des phrases de D_I et leur classification. Il est alors utilisé dans la deuxième étape de notre algorithme, qui consiste à déterminer si une phrase des documents de mise à jour (D_M) doit être classée dans une classe de M_I, ou placée dans une nouvelle classe qui ne contiendrait que des phrases porteuses de nouveauté. L'algorithme de classification que nous utilisons, *fast global k-means*, peut être adapté afin de confronter des éléments à un modèle préétabli, et ce afin de déterminer s'ils doivent intégrer ce modèle. Nous décrivons ici la partie de notre algorithme de résumé de mise à jour dédiée à la détection de nouveauté.

Premièrement, les similarités entre les phrases de D_M et les centres de classe de M_I ainsi qu'entre toutes les phrases de D_M et entre les phrases de D_I sont calculées. Les phrases de D_M sont ensuite ajoutées à M_I, et *fast global k-means* est relancé à partir de la $k_I{}^{\text{ème}}$ itération avec les contraintes suivantes :

– les phrases de D_I ne peuvent pas être dépacées vers un autre cluster, ceci afin de préserver le modèle M_I qui encode les informations initiales. Cela évite également de perturber la portée sémantique des nouvelles classes, *a priori* porteuses de nouveauté, en leur attribuant des éléments initiaux.

– Les barycentres des classes de M_I ne sont pas recalculés. La représentation d'une classe dépend directement de son barycentre ; cela évite donc de modifier la portée sémantique des classes de M_I lorsque des éléments issus de D_M y sont intégrés.

Le principal défaut de cet algorithme détaillé en Figure 5 résidait, dans ses versions précédentes (Bossard, 2011), dans l'absence d'indice de qualité de clustering. Le nombre de classes de mise à jour k_M était alors fixé empiriquement à $\frac{|P_I|}{|P_M|} \times k_I$. Le nombre de classes k_I était lui-même fixé empiriquement à un nombre de phrases dépendant de la taille maximum du résumé. Il était donc possible que même en l'absence de nouveauté, l'algorithme force la création de nouvelles classes, qui auraient alors contenu des informations initiales.

L'utilisation d'un indice de qualité de clustering permet de pallier ce défaut. En effet, le nombre de classes est calculé automatiquement ; le nombre de classes k_M peut donc être égal au nombre de classes k_I, ce qui sous-entend que les documents de mise à jour ne contiennent aucune information nouvelle (ou différente).

4.2. Génération du résumé de mise à jour

Une fois l'algorithme de détection de la nouveauté terminé, k_M nouvelles classes ont été créées. Ce sont ces classes qui vont servir de base à la génération du résumé de mise à jour. Le système génère alors un résumé d'après la méthode exposée dans les sections 3.3 à 3.5, en utilisant uniquement les k_M classes de mise à jour et les phrases qu'elles contiennent.

5. Cas d'étude

Cette section présente quatre cas d'étude ainsi que leur évaluation. Le premier concerne le résumé standard d'actualités. Le second concerne le résumé de mise à jour d'actualités. Ces deux types de résumé ont été évalués sur le corpus français « RPM2 » d'évaluation de résumés.

Les troisième cas d'étude vise à évaluer la capacité de notre système à générer des résumés sur des documents autres que des documents d'actualité, à savoir des aggrégats issus de sources encyclopédiques.

//Classification pour M_I
pour tous les p de P_I, faire
  $cluster[p] \leftarrow C_1$
fin pour
pour i de 1 à $card(P_I)$,faire
  pour n de 1 à i, faire
    $$barycentre[C_n] \leftarrow \frac{\sum_{p_j \in C_n} p_j}{card(C_n)}$$
  fin pour
  pour tous les p dans P_I, faire
    $$cluster[p] \leftarrow \underset{C_m, 1<m<i}{\operatorname{argmax}} \; sim(barycentre[C_m], p)$$
  fin pour
  si $i < card(P_I)$ alors
    $$cluster[\underset{p \in D_I}{\operatorname{argmin}} \; sim(p, barycentre[cluster[p]]) \leftarrow C_(i+1)$$
  fin si
  si $i == 2$ alors
    $indDBmax \leftarrow indiceDavies - Bouldin()$
    $k_I \leftarrow 2$
    enregistrer (cluster); enregistrer (barycentres);
  fin si
  sinon
    si $i > 1$&& $indDBmax < indiceDaviesBouldin()$ alors
      $indDBMax \leftarrow indiceDaviesBouldin()$
      $k_I \leftarrow i$
      enregistrer (clusters); enregistrer (barycentres);
    fin sinon
  fin si
fin pour
//Détection de la nouveauté
charger(clusters); charger(barycentres);
pour tous les p de P_M, faire
  $$cluster[p] \leftarrow \underset{C_i, 1<i<k_I}{\operatorname{argmax}} \; sim(barycentre[C_i], p)$$
fin pour
$indDMax \leftarrow indiceDaviesBouldin$
pour i de k_I à $card(P_M) + k_I$, faire
  pour n de $k_I + 1$ à i, faire
    $$barycentre[C_n] \leftarrow \frac{\sum_{p_j \in C_n} p_j}{card(C_n)}$$
  fin pour
  pour tous les p dans P_M, faire
    $$cluster[p] \leftarrow \underset{C_m, 1<m<k_I}{\operatorname{argmax}} \; sim(barycentre[C_m], p)$$
  fin pour
  si $i < k_M$, alors
    $$cluster[\underset{p_m \in D_I}{\operatorname{argmin}} \; sim(p_m, barycentre[cluster[p]]) \leftarrow C_{i+1}$$
  fin si si $indiceDBMax < indiceDaviesBouldin()$, alors
    $indiceDBMax \leftarrow indiceDaviesBouldin()$
    enregistrer(clusters);
  fin si
fin pour

Figure 5. Algorithme de détection de nouveauté

5.1. Résumé d'actualités

Le résumé d'actualités est la tâche « phare » du résumé automatique. Dans un monde où la quantité d'actualités quotidiennes ainsi que le nombre de sources sont trop importants pour un lecteur, le résumé automatique associé à des techniques de regroupements de documents par thème paraît être une solution efficace pour accéder rapidement aux contenus essentiels de l'information. C'est donc la portée applicative intéressante de ce domaine qui a poussé les organisateurs de campagnes d'évaluation internationales (DUC 2007, TAC 2008, TAC 2009 et TAC 2009) à faire du résumé automatique d'actualités regroupées en thèmes la tâche principale.

Nous présentons ici l'expérience que nous avons menée sur un corpus français développé par le LIA dans le cadre du projet RPM2. Le corpus est téléchargeable gratuitement, sur demande, à l'adresse : http://lia.univ-avignon.fr/fileadmin/documents/rpm2/rpm2_resumes_fr.html.

5.1.1. Description détaillée de la tâche

Le corpus de résumé multidocument RPM2 est divisé en 20 thèmes, et il s'agit de générer un résumé de 100 mots maximum pour chaque thème. Chaque thème est composé de 10 documents, d'origine, de type et de taille variables (Loupy *et al.*, 2010), tous datés de janvier à septembre 2009. Le corpus est proposé sous trois formats différents : texte brut, XML ou HTML. La liste des thèmes est fournie en Tableau 2.

Tableau 2. Liste des thèmes du corpus RPM2

01	Libération d'Ingrid Bétancourt
02	Caisse d'Epargne
03	Crise bancaire
04	Visite du Dalaï Lama en France
05	Fichier Edvige
06	Ouverture des JO de Pékin
07	Jérôme Kerviel
08	Lance Armstrong
09	La loi Leonetti
10	Abandon du petit Mohamed
11	Election d'Obama
12	Licenciement de Patrick Poivre d'Arvor
13	Le temple de Preah Vihear
14	Elections du Secrétaire général du PS
15	Grossesse de Rachida Dati
16	Rachida Dati et les magistrats
17	Réforme du lycée
18	Réforme de l'audiovisuel public
19	Mesures pour la relance de l'économie française
20	Crise au Tibet

5.1.2. Méthode d'évaluation

Le corpus RPM2 comprend quatre résumés de référence par thème, écrits par quatre annotateurs différents. Ces résumés manuels sont d'une longueur de 100 mots,

et peuvent être utilisés pour des évaluations automatiques fondées sur des cooccurrences de n-grammes entre résumés de référence et résumés automatiques, comme celles fournies par le *package* d'évaluation ROUGE [7] (C.-Y. Lin, 2004). Les évaluations ROUGE sont suffisamment corrélées aux notes manuelles pour être considérées comme un indicateur fiable de la qualité d'un résumé automatique (Dang, Owczarzak, 2008). Nous avons donc choisi d'utiliser le *package* ROUGE plutôt que des méthodes semi-automatiques telles que Pyramide (C.-Y. Lin *et al.*, 2006), plus précises, mais dont le temps nécessaire au processus d'évaluation est rédhibitoire dans le cadre d'une démarche expérimentale.

5.1.3. Baselines

Nous avons utilisé deux *baselines* afin de les comparer à notre système de résumé automatique.

La première, que nous nommons Seuil, consiste à calculer les scores *LexRank* (*cf* équation 6) de toutes les phrases. Les phrases sont ensuite extraites dans l'ordre, à condition qu'elles vérifient les deux contraintes présentées dans les équations 10 et 11. Afin de comparer notre système à la meilleure configuration possible de cette *baseline*, le seuil de similarité δ a été calculé expérimentalement pour produire des résumés qui maximisent le score ROUGE-SU4 moyen sur le corpus d'évaluation RPM2 ($\delta = 0.55$) [8].

$$Taille(phrase) < Contrainte_{taille} - \sum_{p \in S} Taille(p) \tag{10}$$

où S est l'ensemble des phrases déjà sélectionnées.

$$\max_{p \in S}(sim(phrase, p)) < \delta \tag{11}$$

où S est l'ensemble des phrases déjà sélectionnées, et δ un seuil de similarité.

La deuxième *baseline* est la fonction MMR avec LexRank comme score de centralité, et la similarité aux autre phrases calculée d'après la mesure cosinus pondérée présentée en équation 2. La meilleure configuration de MMR a été retenue, à savoir un facteur de nouveauté égal à 0,55.

5.1.4. Poids des unités lexicales

Nous n'avions pas décrit le poids associé aux unités lexicales que nous utilisons dans notre système. Celui-ci dépend en effet de l'application visée, et du type de corpus utilisé. Dans le cas du résumé multidocument standard sur le corpus RPM2, nous

7. http://www.berouge.com

8. Nous n'avons pas différencié le corpus d'apprentissage du corpus d'évaluation pour les *baselines*. Cela ne peut que les avantager face à CBSEAS, pour lequel nous n'avons pas cherché à obtenir le score maximal sur le corpus de test.

avons utilisé un *tf.idf* calculé sur les différents thèmes fournis par le corpus. Le *tf.idf* permet de rendre compte de l'importance d'une unité lexicale dans un thème donné relativement aux autres thèmes.

5.1.5. Résultats

Le Tableau 3 présente les scores ROUGE obtenus par les différents systèmes évalués sur le corpus RPM2. Les différentes versions de CBSEAS évalués sont :

– **v1.0** : l'ancienne version de CBSEAS, qui n'inclut pas d'indice de qualité de clustering. Le nombre de classes en entrée de *fast global k-means* dépend du nombre de mots maximum désirés par l'utilisateur ;

– **v2.0** : la nouvelle version de CBSEAS, qui inclut un indice de qualité de clustering ainsi que la similarité au titre ;

– **v2.0 bis** : la nouvelle version de CBSEAS, sans similarité au titre.

L'évaluation de ces différents systèmes nous permet d'évaluer l'apport de l'indice de qualité de clustering, ainsi que celui de la similarité au titre, les deux principales nouveautés apportées au système.

La Figure 6 présente deux résumés différents, l'un généré d'après la méthode MMR, l'autre par CBSEAS v2.0.

Tableau 3. Scores ROUGE des différents systèmes évalués sur les résumés standard du corpus RPM2

	Seuil	MMR	CBSEAS		
			v1.0	v2.0	v2.0 bis
ROUGE1	0.36494	0.37532	0.39666	0.40021	0.39956
ROUGE2	0.12278	0.12222	0.13667	0.13865	0.13802
ROUGE-SU4	0.13848	0.14037	0.15024	0.15367	0.15359

5.2. Résumé de mise à jour d'actualités

Chaque thème du corpus RPM2 comprend deux types différents de documents : les documents initiaux, et les documents de mise à jour. Dans la section précédente, nous n'avons décrit que la partie de l'évaluation dédiée aux documents initiaux, donc au résumé multidocument standard. Cette section présente nos expériences sur le résumé automatique de mise à jour.

5.2.1. Description détaillée de la tâche et évaluation

Les thèmes du corpus RPM2 sont composés de 10 documents initiaux ainsi que de 10 documents de mise à jour. La tâche de résumé de mise à jour consiste à générer, pour chacun des thèmes, un résumé de 100 mots maximum en considérant que le lecteur de ce résumé a déjà lu les documents initiaux. Il souhaite donc ne se voir présenter que des informations nouvelles.

Le laboratoire français antidopage de Châtenay-Malabry (LNDD) avait procédé à une analyse d'échantillons contenant de l'EPO dont six ont été attribués par le journal au coureur américain.
L'Américain Lance Armstrong, septuple vainqueur du Tour de France qui a annoncé dernièrement son retour à la compétition en 2009 dans l'équipe Astana, a repoussé, mercredi 1er octobre, la proposition de l'Agence française de lutte contre le dopage (AFLD) de procéder à une nouvelles analyse des échantillons prélevés pendant le Tour de France 1999 "pour couper court aux rumeurs qui le concernent si elles sont infondées".

Résumé initial 8 CBSEAS (CSR/CSRG) (F-MESURE ROUGE-SU4 F:**0.19185**)

En 1999, l'année de la première victoire d'Armstrong dans le Tour, le laboratoire français antidopage de Châtenay-Malabry (LNDD) avait procédé à une analyse d'échantillons contenant de l'EPO.
" Pour son grand retour à la compétition, le septuple vainqueur du Tour de France, dont le parcours sera dévoilé ce matin, pourrait ainsi ne disputer que le Giro, tandis que Aberto Contador, qui restera chez Astana en 2009, roulerait sur la Grand Boucle.
A l'AFLD, on n est absolument pas d accord avec les excuses données par Armstrong.
Encore aujourd'hui avec le Tour, il y a des doutes.

Résumé initial 8 MMR (CSR/CSRG) (F-MESURE ROUGE-SU4 F:**0.14396**)

Figure 6. Exemple de deux résumés, le premier produit par CBSEAS, le deuxième par MMR

Les résumés sont évalués selon le même protocole que les résumés initiaux (*cf* section 5.1.2).

5.2.2. Baselines

Nous avons implémenté les deux mêmes *baselines* que pour la tâche précédente. Celles-ci sont présentées en section 5.1.3. Pour gérer la nouveauté, nous ajoutons l'ensemble des phrases des documents initiaux à l'ensemble des phrases S déjà sélectionnées, pour la *baseline* MMR comme pour la *baseline* Seuil.

5.2.3. Poids liés aux unités lexicales

Le poids des unités lexicales doit refléter les différences de fréquences entre le corpus initial et le corpus de mise à jour. Chaque terme reçoit donc comme poids sa cross-entropie.

5.2.4. Résultats

Les systèmes dont les résultats sont présentés dans le Tableau 4 sont les mêmes que pour l'évaluation précédente (*cf* section 5.1.5).

Tableau 4. Scores ROUGE des différents systèmes évalués sur les résumés de mise à jour de RPM2

	Seuil	MMR	CBSEAS		
			v1.0	v2.0	v2.0 bis
ROUGE1	0.31497	0.32945	0.36981	0.37632	0.37368
ROUGE2	0.10154	0.10703	0.12453	0.12796	0.12167
ROUGE-SU4	0.11549	0.12413	0.14059	0.14538	0.14017

5.3. Résumé d'aggrégats issus de sources encyclopédiques

Dans cette expérience, nous avons voulu évaluer les résumés générés à partir de sources autres que l'actualité. L'objectif de cette expérience est de générer des fiches synthétiques à propos d'une personnalité donnée. Les informations concernant les personnalités visées sont extraites des données de Wikipédia.

5.3.1. Description du corpus

Nous avons travaillé sur le contenu des pages Wikipédia françaises, téléchargées depuis http://dumps.wikimedia.org/frwiki/latest/frwiki-latest-pages-articles.xml.bz2. Ce fichier contient toutes les pages en texte intégral, structuré par deux catégories différentes de tags XML. La première catégorie délimite les titres de page et les pages elles-mêmes. L'autre catégorie sert à formater les pages de manière structurée (« Infoboxes » ou tableaux), ou de manière semi-structurée (paragraphes, énumérations...).

Les données structurées et semi-structurées sont filtrées afin de travailler à partir des données textuelles uniquement. Notre objectif n'est pas de garder la totalité des informations, mais les informations textuelles uniquement, afin de générer un corpus purement textuel. La syntaxe des tags Wikipédia n'étant liée à aucune DTD, la structure des pages Wikipédia est compliquée à analyser.

Le corpus obtenu est composé pour chaque page d'un ensemble de phrases ou paragraphes. Ces passages sont indexés ; il est alors possible d'extraire tous les passages qui contiennent un ou plusieurs termes visés. Cette indexation des données de Wikipédia nous permet de compléter les informations contenues par une page dédiée à une entité donnée par les informations concernant cette entité disséminées sur d'autres pages Wikipédia. La Figure 7 illustre les données que nous obtenons pour le terme « Tom Cruise ». Cet exemple présente seulement un court extrait de données traitant de « Tom Cruise » issues de pages autres que la page Wikipédia le concernant.

Le corpus ainsi constitué est différent d'un corpus de dépêches : dans un corpus d'actualité, il est possible d'inférer l'importance d'une information d'après le nombre de phrases qui la véhiculent. En revanche, dans notre corpus, un tel postulat n'est pas évident. Ce corpus est donc un moyen d'évaluer l'impact de la classification automatique sur un corpus où la corrélation entre redondance et pertinence ne peut pas être établie de manière certaine.

Stan appelle son père à l'aide et Randy tape à la porte en disant « Mr Cruise, vous ne pouvez pas rester dans le placard ; vous devez en sortir ». Quatre heures plus tard, Cruise est toujours dans le placard et une foule attend pendant que le chef de la police de South Park lui demande à l'aide d'un mégaphone de « sortir du placard ». R.Kelly est sur la scène et chante une brève chanson inspirée de Trapped in the Closet puis menace de « sortir son flingue » si on ne lui dit pas pourquoi Tom Cruise est dans le placard.
De physique jeune et androgyne, Armand a été interprété au cinéma par Antonio Banderas aux côtés de Brad Pitt (Louis) et Tom Cruise (Lestat) dans le film Entretien avec un vampire de Neil Jordan.
Après une rencontre survenue le 30 août 2004 entre Nicolas Sarkozy, ministre français de l'économie et des Finances, et l'acteur américain Tom Cruise, connu pour sa proximité avec la scientologie, M. Brard a mis en cause le ministre, lui reprochant cette rencontre.

Figure 7. Extrait d'un aggrégat encyclopédique concernant « Tom Cruise » : le premier paragrahe provient de la page « Piégé dans le placard » (épisode de South Park », le second de la page « Entretien avec un vampire », et le troisième de la page de « Jean-Pierre Brard »)

Pour notre expérience, qui s'inscrit dans une démarche exploratoire, nous avons généré seulement quatre fichiers traitant de quatre entités différentes : Orange, Tom Cruise, Zinedine Zidane et Norah Jones.

5.3.2. Changements liés à la fonction de score

Le corpus généré en aggrégant du contenu Wikipédia est, dans la forme des phrases, différent des corpus d'actualités utilisés dans la première expérience. Dans ces aggrégats, une forte proportion de phrases débutent par un pronom (*e.g.* « Elle s'est mariée avec Tom Cruise en 2006 ». Ces pronoms doivent soit être résolus, soit supprimés des résumés, car ils peuvent être la source d'erreurs de compréhension. Plutôt que de résoudre les pronoms, nous avons choisi, dans cette étude, de minorer les scores des phrases débutant par un pronom, en les divisant par un facteur 2.

5.3.3. Poids liés aux unités lexicales

La structure des corpus générés ne nous permet pas d'utiliser des poids de type « tf.idf ». Dans cette expérience, nous avons attribué à chaque unité lexicale la fréquence inverse des phrases la contenant, couplée à une stop liste calculée sur un corpus d'actualités de 2Go. Les 150 lemmes les plus fréquents de ce corpus sont filtrés dans les calculs de notre système de résumé, afin d'éviter le bruit qu'ils pourraient engendrer.

5.3.4. Résultats

Le Tableau 5 présente les résultats des deux *baselines* et de notre système de résumé. Contrairement aux deux expériences précédentes, nous n'avons pas évalué le système sans indice de qualité de clustering. Les documents n'ont pas de titre ; le score lié à la similarité avec le titre n'est donc pas utilisé. La Figure 8 présente les quatre résumés générés par CBSEAS.

Résumé sur « Norah Jones »

La chanteuse reçoit cinq Grammy Awards pour cet album, dont celui de la "Meilleure nouvelle artiste".
Norah Jones a vendu plus de 39 millions d'albums dans le monde entier.
La carrière de Norah Jones est lancée en 2002 avec la sortie de son premier album, Come Away with Me, qui se vend à plus de 20 millions d'exemplaires.
Norah et Lee Alexander se sont séparés au début de l'année 2008.
Come Away with Me est le premier album de la chanteuse et pianiste Norah Jones, sorti en 2002.
Feels like Home est le deuxième album de la chanteuse américaine Norah Jones, sorti en 2004.
Norah Jones a eu pendant quatre ans pour compagnon le bassiste Lee Alexander.
Elle se spécialise alors dans le piano jazz et remporte de nombreux prix.
Aux Pays-Bas, ce fut la meilleure vente de l'année.

Résumé sur « Orange »

Tout en restant sur France Télévisions, il rejoint Orange Sports en juin 2008.
Les activités de téléphonie mobile sont gérées par la société Orange France, filiale de Orange SA.
Orange propose aussi en Belgique une Livebox sous la marque Mobistar.
Studio 37 est la filiale cinéma de France Télécom / Orange engagée dans la co-production et l'acquisition de films français et européens.
) pour leur permettre de proposer un service de dictionnaires à leurs utilisateurs.
Orange Réunion est une entreprise française de télécommunications agissant principalement en tant que fournisseur d'accès à Internet et opérateur de téléphonie mobile sur l'île de La Réunion, département d'outre-mer et région ultrapériphérique de l'Union européenne dans le sud-ouest de l'océan Indien.
utilise la marque Orange sous licence pour ses services de téléphonie mobile, sans pour autant appartenir à Orange SA.

Résumé sur « Tom Cruise »

Le film met en scène dans les rôles principaux le couple Tom Cruise et Nicole Kidman.
Il s'est fait connaître à l'international avec la sortie du film Top Gun.
Keri Russell obtient enfin un rôle important au cinéma en 2006 dans Mission : Impossible 3 aux côtés de Tom Cruise.
En 2002, elle tiens un rôle au côté de Tom Cruise dans Minority Report, film réalisé par Steven Spielberg.
Elle a été l'épouse de l'acteur Tom Cruise de 1990 à 2001.
Il recommande l'actrice pour le film Jours de tonnerre, où il tient le rôle principal.
À 15 ans, Tom Cruise ne se prédestinait pas à devenir acteur.
La carrière de Nicole Kidman part en flèche avec Eyes Wide Shut, le dernier film de Stanley Kubrick.
Il reprendra le rôle dans les films 6 et 7.

Résumé sur « Zinedine Zidane »

Ce match est aussi le dernier de Zinedine Zidane au Stade de France.
Ce match fut le dernier match de sa carrière de footballeur.
Zidane conserve cependant son titre de meilleur joueur de la Coupe du monde 2006.
Il devient le premier joueur à porter ce numéro depuis la retraite de Zidane.
Le 9 juillet 2006, lors de la finale de la Coupe du monde 2006, Zinedine Zidane assène un violent coup de tête à l'Italien Marco Materazzi.
zinedine zidane Anecdote à part, au cours de l'année 2008, il fera parti des joueurs invités afin de disputer le match contre la pauvreté qui opposa les amis de Ronaldo à ceux de Zidane.
Selon Le Figaro, Zinedine Zidane gagnerait plus de 300000 euros par match.
La France perd la finale 5-3 aux tirs au but.

Figure 8. Résumés générés par CBSEAS sur le corpus d'aggrégats issus de Wikipédia

Tableau 5. Scores ROUGE des systèmes sur le corpus d'aggrégats issus de Wikipédia

	Seuil	MMR	CBSEAS
ROUGE-1	0.39013	0.39282	0.40860
ROUGE-2	0.10134	0.10654	0.10791
ROUGE-SU4	0.13569	0.13780	0.14649

6. Discussion

Cette section présente l'analyse des résultats obtenus sur les différentes évaluations proposées dans l'article, ainsi qu'une discussion relative à la généricité des systèmes de résumé automatique.

6.1. Conclusions

Les résultats obtenus sur le corpus RPM2 en résumé multidocument standard montrent que notre méthode surpasse largement les deux *baselines*, dont MMR. De plus, le système fondé sur une classification entièrement non supervisée surpasse également les anciennes versions du système, qui ne calculaient pas automatiquement le nombre optimal de classes. La similarité au titre apparaît comme un critère important de sélection de phrases pour générer des résumés automatiques d'actualité.

L'expérience menée sur le résumé de mise à jour a également montré l'efficacité de notre méthode de résumé automatique. Elle vient confirmer les bons résultats obtenus par CBSEAS, sans calcul automatique du nombre de classes, et sur l'anglais, sur les données de TAC 2009 (Bossard, 2011). En revanche, le constat reste le même concernant la difficulté de la tâche de mise à jour : les scores ROUGE des résumés de mise à jour sont moins bons que les scores des résumés initiaux, ce qui suppose une qualité de résumé inférieure. Nous constatons la même augmentation des scores ROUGE des résumés générés en utilisant un indice de qualité de clustering que lors de l'expérience précédente.

La méthode d'évaluation des résumés de mise à jour est critiquable. En effet, elle consiste seulement à calculer le rappel des n-grammes des résumés de référence présents dans les résumés automatiques. Un score ROUGE élevé ne garantit pas que le résumé ne présente pas des informations redondantes par rapport aux documents initiaux. D'autres mesures devraient donc être utilisées afin d'évaluer la non-redondance des résumés de mise à jour.

Les résumés générés à partir des extraits de Wikipédia ne sont pas satisfaisants. Bien que les scores ROUGE des résumés générés par notre système soient, tout comme pour les deux expériences précédentes, plus élevés que ceux des *baselines*, les résumés présentent des défauts importants. Par exemple, on constate que les premières phrases d'une page de Wikipédia, dont la pertinence est importante car elles introduisent l'entité de manière générale, ne sont extraites dans aucun des quatre résumés. Pour Tom Cruise, cela nous aurait permis de savoir que « Thomas Cruise Mapother IV,

Zinedine Yazid Zidane, né le 23 juin 1972 à Marseille, souvent surnommé Zizou, est un footballeur international français.
Il est cité parmi les plus grands joueurs de football de tous les temps et est listé parmi les 125 meilleurs joueurs mondiaux encore vivants en 2004, dans un classement conjoint de Pelé et de la Fédération internationale de football association (FIFA). Sportif préféré des Français en 2006[2], il est classé à trois reprises meilleur joueur mondial de l'année par la FIFA en 1998, 2000 et 2003[3] et ballon d'or en 1998. Il est par deux fois classé second meilleur joueur français de tous les temps par France Football. En 2004, il est élu meilleur joueur européen du demi-siècle par l'UEFA et devance donc des légendes comme Cruijff, Beckenbauer ou encore Platini.
Jouant au poste de milieu offensif, il a été le meneur de jeu de prestigieux clubs européens, comme la Juventus de Turin et le Real Madrid, avec lesquels il a remporté de nombreux titres nationaux et internationaux.

Figure 9. Trois premiers paragraphes de la page Wikipédia de « Zinedine Zidane »

né le 3 juillet 1962 à Syracuse, dans l'État de New York, est un acteur et producteur américain. ». Certes, la lecture (fastidieuse) du résumé nous permet d'inférer le fait que Tom Cruise est un acteur, mais elle ne permet pas de prendre connaissance des informations biographiques essentielles.

Nous avions montré dans (Bossard, 2009) que la structure d'une dépêche de presse pouvait servir à enrichir le calcul de centralité. Il en va de même pour les articles encyclopédiques. Inclure des informations liées à la structure, ou tout du moins à la position des phrases dans les documents, nécessiterait de revoir en profondeur la structure de notre corpus ainsi que la manière d'extraire les données.

Nous avons également remarqué que certains passages des pages dédiées à l'entité visée ne sont pas extraits. Ceci est dû à notre stratégie d'extraction, qui consiste à extraire un passage si et seulement s'il contient l'entité visée. Les trois premiers paragraphes de la page concernant Zinedine Zidane, présentés en Figure 9 ne font pas partie de notre corpus, car ils ne contiennent pas l'entité en question. Extraire systématiquement l'intégralité de la page dédiée à l'entité visée permettrait de remédier à ce problème.

6.2. Perspectives

Les performances relativement bonnes de notre système sur le résumé d'actualité nous ont amenés à soulever le problème de sa généricité : quelles en sont les limites applicatives ? L'article a ainsi présenté une expérience menée sur des aggrégats issus de Wikipédia, concernant une entité donnée. Si les performances du système en termes de score ROUGE-1 sont acceptables, les résumés ne permettent pas de prendre directement connaissance des informations essentielles concernant l'entité visée. Dès lors, se pose la question des limites des systèmes purement statistiques, et de leur avantage supposé face aux systèmes symboliques concernant la généricité.

De nombreux travaux ont montré que des traitements spécifiques étaient nécessaires à la génération de résumés sur des domaines spécifiques, comme les textes juridiques (Farzindar, Lapalme, 2005) ou les fils d'e-mails (Wan, McKeown, 2004). Nos prochains travaux s'attacheront donc à étudier les domaines sur lesquels les approches statistiques actuelles produisent des résultats statisfaisants, et à voir si des traitements simples peuvent les améliorer.

Nous nous attacherons également à corriger les défauts d'extraction du corpus Wikipédia, afin de pouvoir disposer d'un corpus d'une qualité suffisante à des expérimentations plus poussées.

Bibliographie

Adda G., Mariani J., Paroubek P., Rajman M., Lecomte J. (1999). Métrique et premiers résultats de l'évaluation grace des étiquetteurs morpho-syntaxiques pour le français. *TALN 1999*, p. 15-24.

Barzilay R., Elhadad N., McKeown K. (2002). Inferring Strategies for Sentence Ordering in Multidocument News Summarization. *J. Artif. Intell. Res. (JAIR)*, vol. 17, p. 35-55.

Bossard A. (2009). Une approche mixte – statistique et structurelle – pour le résumé automatique de dépêches. In *Actes de TALN 2009.* Senlis, France.

Bossard A. (2010). *Contribution au résumé automatique multidocument*. Unpublished doctoral dissertation, Université Paris 13, Laboratoire d'informatique de Paris-Nord et CNRS UMR 7030, Villetaneuse, France.

Bossard A. (2011). Generating update summaries : Using an unsupervized clustering algorithm to cluster sentences. In *Multi-source, multilingual information extraction and summarization.* Springer.

Bossard A., Généreux M., Poibeau T. (2010). Résumé automatique de textes d'opinion. *TAL*, vol. 51, n° 3, p. 47–73. http://www.atala.org/IMG/pdf/2-Bossard-TAL51-3.pdf

Bossard A., Guimier De Neef E. (2011a). Etude de l'impact du regroupement automatique de phrases sur un système de résumé multi-documents. In *Actes de la 8e conférence en recherche d'information et applications (coria).* Avignon, France.

Bossard A., Guimier De Neef E. (2011b). Ordonner un résumé automatique multidocument fondé sur une classification des phrases en sous-thèmes - application au résumé de dépêches. In *18e conférence sur le Traitement Automatique du Langage Naturel (taln'11).* Montpellier, France.

Bossard A., Rodrigues C. (2011). Combining a multi-document update summarization system – CBSEAS – with a genetic algorithm. In I. Hatzilygeroudis, J. Prentzas (Eds.), *Combinations of intelligent methods and applications.* Springer.

Boudin F., Torres-Moreno J.-M., El-Bèze M. (2008). A scalable MMR approach to sentence scoring for multi-document update summarization. In *Proceedings of the 2008 coling conference*, p. 21-24. Manchester, UK.

Carbonell J., Goldstein J. (1998). The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *Sigir '98: Proceedings of the 21st annual international acm sigir conference*, p. 335–336. New York, NY, USA, ACM.

Dang H. T., Owczarzak K. (2008). Overview of the TAC 2008 update summarization task. In *Notebook papers and results of TAC 2008*, p. 10-23. Gaithersburg, Maryland, USA.

Darling W., Song F. (2011). Pathsum: A summarization framework based on hierarchical topics. In *Proceedings of text summarization 2011 workshop (ts'11) – canadian ia 2011.* St John's, Newfoundland, Canada.

Davies D. L., Bouldin D. W. (1979, April 27). A Cluster Separation Measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PAMI-1, n° 2, p. 224–227.

Dunn J. C. (1973). A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters. *Journal of Cybernetics*, vol. 3, n° 3, p. 32–57. http://dx.doi.org/10.1080/01969727308546046

Edmundson H. P., Wyllys R. E. (1961). Automatic abstracting and indexing—survey and recommendations. *Commun. ACM*, vol. 4, n° 5, p. 226–234.

Erkan G., Radev D. R. (2004). Lexrank: Graph-based centrality as salience in text summarization. *Journal of Artificial Intelligence Research (JAIR)*, vol. 22.

Farzindar A., Lapalme G. (2005, jun). Production automatique du résumé de textes juridiques: évaluation de qualité et d'acceptabilité. In *Taln 2005*, vol. 1, p. 183-192. Dourdan, France.

Flesch R. (1948). A new readability yardstick. *Journal of applied psychology*, vol. 32, n° 3, p. 221–233.

Galanis D., Malakasiotis P. (2008). Aueb at TAC 2008. In *Notebook papers and results of TAC 2008.* Gaithersburg, Maryland, USA.

Guérif S. (2006). *Réduction de dimension en apprentissage numérique non supervisé*. Unpublished doctoral dissertation, Université Paris 13, Laboratoire d'informatique de Paris-Nord et CNRS UMR 7030, Villetaneuse, France.

He T., Chen J., Gui Z., Li F. (2008). Ccnu at TAC 2008: Proceeding on using semantic method for automated summarization yield. In *Notebook papers and results of TAC 2008.* Gaithersburg, Maryland, USA.

Heinecke J., Smits G., Chardenon C., Guimier De Neef E., Maillebuau E., Boualem M. (2008). Tilt : plate-forme pour le traitement automatique des langues naturelles. *TAL*, vol. 2, n° 49.

Kupiec J., Pedersen J., Chen F. (1995). A trainable document summarizer. In *Sigir '95: Proceedings of the 18th annual international acm sigir conference on research and development in information retrieval*, p. 68–73. New York, NY, USA, ACM.

Lin C.-Y. (2004). Rouge: a package for automatic evaluation of summaries. In *Proceedings of the workshop on text summarization branches out (was 2004).* Barcelona, Spain.

Lin C.-Y., Cao G., Gao J., Nie J.-Y. (2006). An information-theoretic approach to automatic evaluation of summaries. In *HLT NACACL' 2006*, p. 463–470. Morristown, NJ, USA, ACL.

Lin Z., Hoang H. H., Qiu L., Ye S., Kan M.-Y. (2008). NUS at TAC 2008: Augmenting timestamped Graphs with event information and selectively expanding opinion contexts. In *Proceedings of TAC 2008 workshop on automatic summarization.*

López-Escobar S., Carrasco-Ochoa J. A., Trinidad J. F. M. (2006). Fast global *k*-means with similarity functions algorithm. In *Proceedings of the intelligent data engineering and automated learning - ideal 2006, 7th international conference*, p. 512-521.

Loupy C. de, Guégan M., Ayache C., Seng S., Torres Moreno J. M. (2010). A french human reference corpus for multi-document summarization and sentence compression. In *Proceedings of the 7th language resources and evaluation conference.* Valletta, Malte.

Luhn H. (1958). The automatic creation of literature abstracts. *IBM Journal*, vol. 2, n° 2, p. 159-165.

MacQueen J. B. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of 5th berkeley symposium on mathematical statistics and probability*, p. 281–297. University of California Press.

Marcu D. (1998). *Improving summarization through rhetorical parsing tuning.*

Radev D., Winkel A., Topper M. (2002). Multi document centroid-based text summarization. In *Proceedings of the acl 2002 demo session.* Philadelphia, PA, USA.

Rousseeuw P. (1987, November). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, vol. 20, n° 1, p. 53–65. http://dx.doi.org/10.1016/0377-0427(87)90125-7

Rutledge L., Bailey B., Ossenbruggen J. V., Hardman L., Geurts J. (2000). Generating presentation constraints from rhetorical structure. In *Proceedings of the 11th ACM conference on hypertext and hypermedia*, p. pages. ACM.

Saggion H., Gaizauskas R. (2004). Multi-document summarization by cluster/profile relevance and redundancy removal. In *Proceedings of the Document Understanding Conference 2004.*

Varma V., Bysani P., Reddy K., Bharat V., Kovelamudi S., Maganti N. (2009). IIT hyderabad at TAC 2009. In *Notebook papers and results of TAC 2009.* Gaithersburg, Maryland, USA.

Wan S., McKeown K. (2004). Generating overview summaries of ongoing email thread discussions. In *Proceedings of the international conference on computational linguistics (COLING)*, p. 549–555. Geneva, Switzerland.

Wang Y.-W. (2002). *Sentence Ordering for Multi-Document Summarization in Response to Multiple queries.*