
Etude de l'impact du regroupement automatique de phrases sur un système de résumé multi-documents

Aurélien Bossard
Emilie Guimier De Neef

Orange Labs
2, av Pierre Marzin
22307 Lannion CEDEX

RÉSUMÉ. Dans cet article, nous comparons les résultats produits par différentes approches de résumé multi-documents. Nous opposons deux approches classiques à la nôtre qui place la modélisation de la diversité informationnelle du corpus au centre du processus. Nous évaluons également l'impact de différentes mesures de similarité entre phrases. Les expériences, menées sur le corpus RPM2, montrent qu'un regroupement des phrases en classes sémantiques améliore la qualité des résumés.

ABSTRACT. This paper introduces the experiences we led in order to evaluate the impact of a sentence unsupervised clustering algorithm on a multi-document summarization system. We compared this system to two others which implement classic summarization methods. We also evaluated the impact of different sentence similarity measures on the quality of the summaries. The experiences show that a clustering prior to sentence selection does improve the quality of the summaries.

MOTS-CLÉS : Résumé automatique, classification automatique, mesures de similarité textuelle

KEYWORDS: Automatic summarization, automatic clustering, textual similarity metrics

1. Introduction

Les systèmes de résumé automatique font appel à des techniques de plus en plus variées. Les campagnes d'évaluation récentes (TAC 2008 et 2009) ont montré que des systèmes de résumé automatique utilisant des techniques de classification de phrases en classes thématiques côtoient dans le haut du tableau d'autres systèmes qui n'en utilisent pas. Nous avons voulu étudier l'impact, à technologie équivalente, que peut avoir sur un système de résumé automatique multi-documents, le regroupement automatique des phrases traitant d'un même sous-thème préalablement à la phase d'extraction de phrases. Nous présentons dans cet article l'évaluation que nous avons menée sur un corpus dédié au résumé multi-documents de textes français et développé dans le cadre du projet RPM2¹ par le Laboratoire informatique d'Avignon et Syllabs.

Dans un premier temps, nous présentons un bref aperçu des techniques de résumé automatique. Nous présentons ensuite notre système, CBSEAS, et les deux baselines qui servent à évaluer l'intérêt de la classification, avant de procéder à la présentation et à l'analyse des résultats obtenus.

2. Etat de l'art

Nous présentons ici un état de l'art non exhaustif qui vise à une bonne compréhension des techniques utilisées lors de nos expériences. Pour un état de l'art plus complet, vous pouvez vous référer à (Bossard *et al.*, 2008).

(Radev *et al.*, 2002) ont défini la centralité d'une phrase comme sa similarité au contenu discriminant des documents à résumer. Les auteurs font intervenir le résumé automatique dans le cadre d'une collection documentaire dont les documents ont été classifiés automatiquement selon leur sujet. Ils font émerger pour chaque sujet, un vecteur de mots discriminants, appelé « centroïde » en se fondant sur leur *tf.idf* dans la collection documentaire. La similarité entre une phrase et le centroïde est calculée d'après la mesure « cosinus » (*cf* Figure 1).

$$\text{cosinus}_{tf.idf}(P, C) = \frac{\sum_{m \in P, C} tf_{m,P} tf_{m,C} (idf_m)^2}{\sqrt{\sum_{m \in P} (tf_{m,P} idf_m)^2} \sqrt{\sum_{m \in C} (tf_{m,C} idf_m)^2}}$$

Figure 1. La similarité cosinus entre une phrase P et un centroïde C

(Erkan *et al.*, 2004) ont proposé une autre approche, LexRank, qui a l'avantage de s'affranchir plus aisément du calcul du *tf.idf* sur une collection documentaire, et donc d'être applicable indépendamment du cadre dans lequel les résumés doivent être produits. Cette approche se fonde sur la notion de « prestige », et reprend en l'adaptant l'algorithme « PageRank » afin de calculer l'importance d'une phrase vis-à-vis des documents à résumer. Plus une phrase sera similaire à des phrases au prestige important, plus son prestige sera grand. La formule du calcul incrémental de LexRank décrite en

1. Projet RPM2 : <http://labs.sinequa.com/rpm2/>

Figure 2 est appliquée sur le graphe des documents à résumer, où les nœuds sont les phrases et les arêtes leurs similarités, jusqu'à convergence.

$$prest(u) = \frac{d}{N} + (1 - d) \sum_{v \in adj[u]} \frac{\cosinus_{tf.idf}(u, v) \times prest(v)}{\sum_{z \in adj[v]} \cosinus_{tf.idf}(z, v)}, \text{ où :}$$

N est le nombre total de nœuds dans le graphe,
 $adj[u]$ est l'ensemble des nœuds connexes à u .

Figure 2. La formule de calcul de LexRank

Dans ces deux types d'approche, la diversité informationnelle du résumé produit peut être vue comme un compromis entre centralité et non redondance. Les différentes composantes informationnelles du groupe de documents à modéliser ne sont pas identifiées en tant que telles mais résultent du filtrage de la redondance. Nous proposons une approche différente dans laquelle nous cherchons à modéliser cette diversité au travers d'un regroupement des phrases en classes informationnelles. La non redondance est chez nous une conséquence de cette partition de l'information.

3. Le système CBSEAS

Tout comme LexRank, CBSEAS (Bossard *et al.*, 2008) construit un modèle graphique des documents à résumer, où les nœuds sont les phrases, et les arêtes les similarités entre les nœuds. Cette similarité est calculée par une mesure qui travaille sur des paires de lemmes, mais avec une notion de paire souple. En effet, deux lemmes sont appariés s'ils sont séparés par au plus 4 lemmes. Cet appariement pallie l'absence de modélisation sémantique de l'énoncé et permet de récupérer des éléments qui cooccurrent mais sont séparés par des insertions ou des inversions, phénomènes très fréquents dans les corpus de presse. Cette mesure de similarité, que nous appelons $\cosinus_{tf.idf}$ 2-gramme S4, entre deux vecteurs de paires de lemmes X et Y, est présentée en Fig. 3.

$$co2gS4_{tf.idf}(X, Y) = \frac{\sum_{(l_1, l_2) \in X, Y} (tf.idf_{l_1} \times tf.idf_{l_2})^2}{\sqrt{\sum_{(l_1, l_2) \in X} (tf.idf_{l_1} \times tf.idf_{l_2})^2} \times \sqrt{\sum_{(l_1, l_2) \in Y} (tf.idf_{l_1} \times tf.idf_{l_2})^2}}$$

Figure 3. Formule de la similarité $\cosinus_{tf.idf}$ 2-gramme S4

Le graphe est ensuite partitionné par un algorithme de classification non supervisée, *fast global k-means*, présenté dans (López-Escobar *et al.*, 2006). Cette étape permet de faire émerger n classes thématiques dont les phrases sont regroupées selon leurs similarités. Un *tf.icf* est alors calculé pour tous les mots de chaque classe ; il est calculé de la même manière qu'un *tf.idf*, mais en prenant les classes thématiques en lieu et place des documents.

La dernière étape consiste à extraire incrémentalement une phrase par classe thématique, qui valide la contrainte de l'Équation 1, et qui maximise le score de l'Équa-

$$Taille(phrase) < Taille_{maximum} - \sum_{p \in S} Taille(p) \quad [1]$$

où S est l'ensemble des phrases déjà sélectionnées.

$$score(p) = \alpha \times score_{centralite_{globale}}(p) + \beta \times score_{centralite_{locale}}(p) \quad [2]$$

tion 2. Cette étape est itérée jusqu'à ce qu'aucune classe thématique dont une phrase n'a pas déjà été extraite ne contienne au moins une phrase qui valide la contrainte de l'Équation 1. Le score de centralité globale reflète la centralité d'une phrase par rapport au contenu de l'ensemble des documents, tandis que le score de centralité locale reflète la centralité d'une phrase vis-à-vis du contenu du groupe de phrases dans laquelle elle a été classée. La centralité est calculée selon soit un score centroïde, soit un score *LexRank*. Ce n'est pas tant l'impact de ces scores que nous cherchons à évaluer, mais celui du regroupement en classes thématiques préalable à la sélection de phrases. Cette méthode de sélection est notée « *selec1* » dans les résultats.

Nous avons également utilisé une variante de cette méthode de sélection, « *selec2* ». En effet, la stratégie précédente risque de pénaliser des phrases dont un mot est présent dans plusieurs classes, et a donc un *tf.icf* faible. Après avoir extrait une phrase d'une classe, cette classe est supprimée, et les *tf.icf* sont recalculés avant de sélectionner la phrase suivante. Ainsi, les mots d'une classe thématique déjà extraite qui co-occurrent avec d'autres classes thématiques auront un impact plus important dans le calcul de la similarité locale, dû à l'augmentation de leur *tf.icf*. Une autre solution serait de s'affranchir du *tf.icf* et de se concentrer sur la fréquence des mots, mais cela nécessite de mettre au point un antidiCTIONNAIRE.

3.1. Pré-traitements

En amont de CBSEAS, les documents à résumer sont lemmatisés par TiLT-Lemmatiseur (Heinecke *et al.*, 2008) qui assure la désambiguïstation morpho-syntaxique des segments de la phrase au moyen d'une grammaire hors contexte de chunking. Pour le français, la grammaire comporte plus de 2000 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 TreeTagger² et avait été évalué dans le cadre de la campagne GRACE (Adda *et al.*, 1999) avec une précision supérieure à 95%.

Nous nous plaçons dans un cas d'utilisation où un utilisateur cherche à obtenir le résumé d'un groupe de documents homogènes en contenu, extrait d'une collection documentaire plus large. Il est donc possible d'obtenir une valeur de discrimination d'un terme pour un groupe de documents donné, grâce à un calcul de type *tf.idf*. Le

2. <http://www.ims.uni-stuttgart.de/~schmid/french-tagset.html>

$tf.idf$ de chaque lemme identifié par Tilt est donc calculé avant l'application de la méthode CBSEAS.

4. Évaluation

Nous avons voulu évaluer notre travail sur la langue française. Un corpus d'évaluation de résumé multi-documents et de compression de phrases a été développé dans le cadre du projet RPM2 (de Loupy *et al.*, 2010). La partie évaluation de résumé multi-documents comporte 400 documents divisés en 20 thèmes. Chacun des thèmes comporte deux jeux de 10 documents : un jeu de documents initiaux, et un jeu de documents de mise à jour, visant à évaluer des systèmes de résumé incrémental. Nous nous sommes concentré sur l'évaluation des résumés standards, et nous nous sommes fondé uniquement sur les documents initiaux pour réaliser les résumés.

Afin d'évaluer l'impact que peut avoir la phase de classification préalable à la sélection de phrases, nous avons implémenté un système *baseline*, semblable à celui utilisé dans (Radev *et al.*, 2004) et (Erkan *et al.*, 2004).

4.1. Baseline

Cette *baseline* calcule les scores de chaque phrase dans une première étape. Les phrases sont ensuite ordonnées selon leur score. Les phrases sont enfin extraites dans l'ordre, à la condition qu'elles vérifient les deux contraintes présentées dans les Équations 3 et 4. Afin de comparer notre système à la meilleure configuration possible de cette *baseline*, le seuil de similarité δ a été calculé expérimentalement pour produire le jeu de résumés qui maximise le score ROUGE-SU4 sur le corpus d'évaluation RPM2 ($\delta = 0.55$)³.

$$Taille(phrase) < Taille_{maximum} - \sum_{p \in S} Taille(p) \quad [3]$$

où S est l'ensemble des phrases déjà sélectionnées.

$$\max_{s \in S} (sim(phrase, s)) < \delta \quad [4]$$

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

4.2. Mesures d'évaluation

Il existe plusieurs types de méthodes pour évaluer un résumé. Le premier type de méthode, utilisé entre autres méthodes lors des campagnes TAC notamment (Dang *et*

3. 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.

al., 2008), consiste à demander à un juge d’attribuer une note qui reflète la qualité du résumé. Cette méthode d’évaluation « subjective » présente le défaut d’être particulièrement sensible à la qualité de présentation d’un résumé, et donc d’évaluer moins franchement la quantité d’information présentée dans ce même résumé.

La méthode Pyramide (Lin *et al.*, 2006) est une méthode d’évaluation manuelle de l’informativité des résumés. Elle permet certes d’évaluer de manière extrêmement précise la quantité et la qualité d’information présente dans un résumé, mais nécessite des moyens humains considérables qui la rendent difficilement exploitable dans un cadre expérimental. Pour cette raison, nous nous sommes tourné vers des mesures automatiques de l’informativité. Plusieurs méthodes existent, comme les méthodes ROSE (Conroy *et al.*, 2008) et ROUGE (Lin, 2004), fondées sur de la comparaison de n-grammes entre résumés de référence et résumés automatiques, ou la méthode des *Basic Elements* (Hovy *et al.*, 2006), qui présente l’inconvénient d’être dépendante de traitements linguistiques poussés, et donc difficilement portable vers une autre langue que l’anglais, pour laquelle elle a été conçue. Nous avons choisi de travailler avec les métriques ROUGE, qui sont les plus couramment utilisées lors de campagnes d’évaluation, et fournissent une valeur de comparaison avec des systèmes existants.

4.3. Résultats

Nous présentons ici les résultats que nous avons obtenus lors de trois protocoles d’évaluation différents. Le premier visait à déterminer l’impact du regroupement de phrases non-supervisé sur le résumé automatique. Le système CBSEAS est donc comparé aux implémentations de nos deux *Baselines* pour les mesures de centralité « Centroïde », « LexRank ». Le Tableau 1 présente les scores ROUGE de CBSEAS de la *Baseline* présentée en Section 4.1.

	Centroïde			LexRank		
	Baseline	CBSEAS		Baseline	CBSEAS	
		selec1	selec2		selec1	selec2
ROUGE1	0.35316	0.39666	0.39304	0.36494	0.38501	0.39296
ROUGE2	0.11342	0.13202	0.12815	0.12278	0.13398	0.13667
ROUGE-SU4	0.13495	0.15104	0.14741	0.13848	0.14698	0.15024

Tableau 1. Scores ROUGE des différents systèmes évalués

4.4. Discussion

Le Tableau 1 montre que notre système surpasse les deux *baselines* avec lesquelles il a été confronté. Les différences avec la *Baseline 2*, qui implémente la méthode MMR, oscillent entre 7% et 9% pour le score ROUGE-SU4, en faveur de CBSEAS. Des deux stratégies de sélection que nous avons implémentées, la première semble la plus adaptée à la mesure de centralité « Centroïde », et la seconde à la mesure *LexRank*. Nous expliquons cela par le fait que les centroïdes étant limités en taille (pour les classes thématiques, nous avons choisi un centroïde de dix mots), ceux-ci finissent

Résumé et classification de phrases

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

par contenir majoritairement les mots les plus fréquents dans le jeu de documents considéré, plutôt que les mots les plus discriminants de leur classe thématique. La mesure de centralité *LexRank* est moins sensible à ce phénomène, puisqu'elle prend en compte la totalité du vocabulaire des phrases.

Les mesures ROUGE permettent une mesure de la quantité d'information présente dans les résumés. Si l'on compare les résumés de la Figure 4 ci-dessous, on note que le refus de Lance Armstrong de se soumettre aux analyses est une information manquante dans le second résumé qui intervient probablement dans la différence des scores obtenus. Néanmoins, ces mesures n'évaluent ni l'intérêt du résumé pour l'utilisateur, ni sa lisibilité, ni leur qualité linguistique (redondance, cohérence interne et rhétorique...). Typiquement, dans le résumé (1), on peut identifier un problème de cohérence puisque la périphrase 'coureur américain' fait référence à Lance Armstrong, nommé dans la seconde. Dans ce cas, l'ordonnancement des phrases est en cause et constitue l'un des axes pour la poursuite de ce travail. Des travaux sur la compression des énoncés sont également à l'étude pour traquer la redondance et maximiser l'information utile au sein des résumés.

5. Conclusion

Dans cet article, nous avons présenté les expériences que nous avons menées autour du système de résumé automatique multi-documents CBSEAS, qui gère la diversité et la centralité informationnelles en se servant d'un modèle établi d'après une classification automatique des phrases à résumer. Ces expériences nous ont permis de montrer l'efficacité de la classification de phrases. À méthode de mesure de la centralité équivalente, les systèmes utilisant la classification surpassent systématiquement ceux qui ne l'utilisent pas.

Cependant, ces travaux ont été évalués par les mesures automatiques ROUGE, qui, si elles fournissent une approximation satisfaisante de l'évaluation de la qualité

e soumission à *CORIA 2011*

d'un résumé, ne sont pas suffisantes. Ces résultats doivent donc être confirmés par des mesures d'évaluation manuelle, ce qui sera l'objet de prochains travaux.

Enfin, nous n'avons présenté ici que la partie du système CBSEAS qui vise à extraire les phrases les plus pertinentes. La question de l'assemblage de ces phrases se pose, afin de rendre les résumés compréhensibles par leur lecteur. Le modèle de classification de phrases peut servir de base à un réordonnement des phrases. Nous évaluerons différentes méthodes de réordonnement dans des travaux ultérieurs.

6. Bibliographie

- Adda G., Mariani J., Paroubek P., Rajman M., Lecomte J., « Métrique et premiers résultats de l'évaluation GRACE des étiquetteurs morpho-syntaxiques pour le français », *TALN 1999*, p. 15-24, 1999.
- Bossard A., Généreux M., Poibeau T., « Description of the LIPN Systems at TAC2008 : Summarizing Information and Opinions », *TAC 2008, Workshop on Summarization Track*, NIST, Gaithersburg, Maryland USA, 2008.
- Conroy J. M., Dang H. T., « Mind the gap : dangers of divorcing evaluations of summary content from linguistic quality », *COLING' 2008*, ACL, Morristown, NJ, USA, p. 145-152, 2008.
- Dang H. T., Owczarzak K., « Overview of the TAC 2008 Update Summarization Task (DRAFT) », *Notebook papers and results of TAC 2008*, Gaithersburg, MD, USA, p. 10-23, 2008.
- de Loupy C., Guégan M., Ayache C., Seng S., Moreno J.-M. T., « A French Human Reference Corpus for Multi-Document Summarization and Sentence Compression », *LREC'10*, 2010.
- Erkan G., Radev D. R., « LexRank : Graph-based Centrality as Saliency in Text Summarization », *Journal of Artificial Intelligence Research (JAIR)*, 2004.
- Heinecke J., Smits G., Chardenon C., Guimier De Neef E., Maillebau E., Boualem M., « TILT : plate-forme pour le traitement automatique des langues naturelles », *TAL*, 2008.
- Hovy E., Lin C.-Y., Zhou L., Fukumoto J., « Automated Summarization Evaluation with Basic Elements », *LREC' 2006*, 2006.
- Lin C.-Y., « ROUGE : a Package for Automatic Evaluation of Summaries », *Workshop on Text Summarization Branches Out (WAS 2004)*, Barcelona, Spain, 2004.
- Lin C.-Y., Cao G., Gao J., Nie J.-Y., « An information-theoretic approach to automatic evaluation of summaries », *HLT NACACL' 2006*, ACL, Morristown, NJ, USA, p. 463-470, 2006.
- López-Escobar S., Carrasco-Ochoa J. A., Trinidad J. F. M., « Fast Global k -Means with Similarity Functions Algorithm », in , E. Corchado, , H. Yin, , V. J. Botti, , C. Fyfe (eds), *IDEAL*, vol. 4224 of *Lecture Notes in Computer Science*, Springer, p. 512-521, 2006.
- Radev D., Allison T., Blair-Goldensohn S., Blitzer J., Çelebi A., Dimitrov S., Drabek E., Hakim A., Lam W., Liu D., Otterbacher J., Qi H., Saggion H., Teufel S., Topper M., Winkel A., Zhu Z., « MEAD - a platform for multidocument multilingual text summarization », *LREC' 2004*, Lisbon, Portugal, May, 2004.
- Radev D., Winkel A., « Multi Document Centroid-based Text Summarization », *ACL' 2002*, 2002.