TALN 2011, Montpellier, 27 juin – 1er juillet 2011

Ordonner un résumé automatique multi-documents fondé sur une classification des phrases en classes lexicales

Aurélien Bossard Émilie Guimier De Neef
Orange Labs
2 av. Pierre Marzin
22300 Lannion, France
prenom.nom@orange-ftgroup.com

Résumé. Nous présentons différentes méthodes de réordonnancement de phrases pour le résumé automatique fondé sur une classification des phrases à résumer en classes thématiques. Nous comparons ces méthodes à deux *baselines* : ordonnancement des phrases selon leur pertinence et ordonnancement selon la date et la position dans le document d'origine. Nous avons fait évaluer les résumés obtenus sur le corpus RPM2 par 4 annotateurs et présentons les résultats.

Abstract. We present several sentence ordering methods for automatic summarization which are specific to multi-document summarizers, based on sentences subtopic clustering. These methods are compared to two baselines : sentence ordering according to pertinence and according to publication date and inner document position. The resulting summaries on RPM2 corpus have been evaluated by four judges.

Mots-clés : Résumé automatique, ordonnancement de phrases.

Keywords: Automatic summarization, sentence ordering.

1 Introduction

Les systèmes de résumé automatique multi-documents par extraction fondés sur une classification préalable des phrases en classes lexicales (CL) ont récemment prouvé leur efficacité lors des campagnes d'évaluation internationales TAC [1]. Ces systèmes offrent une modélisation du corpus à résumer différente de celles traditionnellement utilisées . La redondance dans les corpus multi-documents est plus importante que dans des documents simples. Une telle modélisation le prend en compte, et s'appuie sur la redondance pour cibler l'information pertinente tout en évitant la répétition d'éléments d'information dans le résumé. Nous avons montré que cette modélisation pouvait être utile afin d'affiner la sélection de phrases (Bossard & De Neef, 2011).

La problématique du réordonnancement de phrases pour le résumé multi-documents est plus complexe que pour le résumé mono-document. En effet, dans un document unique, les phrases sont toutes issues de la même structure discursive, et peuvent être restituées dans l'ordre du document source. En revanche, en multi-documents, les phrases peuvent être extraites de documents épars, et une structure discursive doit être recomposée. Nous montrons ici que la modélisation en classes lexicales peut également servir à réordonner les phrases du résumé, et présentons une évaluation de notre méthode de réordonnancement. Dans un premier temps, nous dressons un état de l'art de l'ordonnancement de phrases pour le résumé automatique. Cet état de l'art vise à renseigner le lecteur sur les différentes stratégies de réordonnancement de résumé, non sur les méthodes de résumé en soi. Pour plus de renseignements sur les méthodes de résumé automatique, le lecteur pourra se référer à (Mani, 1999) ou à (Das & Martins, 2007). Dans un second temps, nous présentons notre méthode d'ordonnancement et son évaluation réalisée sur le corpus RPM2 (de Loupy *et al.*, 2010). Enfin, nous discutons les résultats obtenus

1. *Text Analysis Conference*, organisée par le *National Institute of Science and Technology* : `http://www.nist.gov/tac`

2 État de l'art

À côté de modélisations chronologiques ou rhétoriques, peu robustes et non génériques (Lin *et al.*, 2008; Rutledge *et al.*, 2000), des auteurs ont privilégié des approches fondées uniquement sur des indices lexicaux, tentant de maximiser le vocabulaire commun de deux phrases voisines dans le résumé (Wang, 2002), ou fondées sur la position des phrases dans leur document d'origine et la date de ce dernier (Saggion & Gaizauskas, 2004). Ces approches ne garantissent cependant pas la cohérence logique ni chronologique du résumé. La position d'une phrase dans un document permet difficilement d'en déduire la position dans un texte différent.

Pour répondre à ce problème, Regina Barzilay *et al.* (2002) génèrent des résumés fondés sur la détection de paraphrases. Une phrase est extraite de chacun des groupes paraphrastiques, et ces mêmes groupes sont utilisés afin de déterminer un ordre entre les phrases. Ainsi, ce ne sont plus les caractéristiques d'une phrase isolée qui déterminent sa position dans le résumé, mais les caractéristiques des phrases d'un groupe paraphrastique. L'algorithme MO – *Majority Ordering* – construit un graphe orienté des groupes paraphrastiques. Le poids du lien d'un groupe G_1 vers un groupe G_2 est égal au nombre de documents dans lequel une phrase classée dans G_1 précède une phrase classée dans G_2. Ordonner les phrases du résumé revient alors à ordonner les groupes paraphrastiques dont elles sont issues.

Notre approche est semblable à ce dernier algorithme. Cependant, nous ne fondons pas le résumé automatique sur la détection de paraphrases au sens strict – (Regina Barzilay *et al.*, 2002) détectent les paraphrases par comparaison d'arbres syntaxiques et ajout d'informations sémantiques – mais sur la classification lexicale des phrases. Par conséquent, les classes que nous utilisons sont plus étendues, et les documents comportant plusieurs phrases d'au moins deux CL sont plus nombreux, ce qui oblige à revoir les approches de création du graphe et du parcours de celui-ci.

3 Notre méthode d'ordonnancement

3.1 Le résumé par classification en classes lexicales – CBSEAS

Nous fondons notre approche de résumé automatique sur le système CBSEAS, présenté en détails dans (Bossard & De Neef, 2011). Celui-ci procède en deux étapes : regrouper les phrases en CL, puis sélectionner les phrases à extraire, à raison d'une phrase par classe au maximum. Dans le système utilisé pour l'article, le nombre de classes est fixé à 8 [2]. L'intuition derrière ce système est que la classification des phrases permet d'une part d'éviter la redondance dans les résumés générés, d'autre part de disposer de deux critères d'extraction : la pertinence d'une phrase vis-à-vis du contenu global des documents (centralité globale) et la pertinence d'une phrase vis-à-vis de sa CL (centralité locale). La campagne d'évaluation TAC 2009 témoigne de l'efficacité de ce système de résumé pour ce qui est de l'extraction d'informations essentielles (Bossard & Poibeau, 2009). Ces résultats sont corroborés par une expérience sur un corpus français (Bossard & De Neef, 2011) . En revanche, les campagnes TAC ont révélé des problèmes de lisibilité chez les utilisateurs. Nous supposons que ceci est dû pour une grande partie au manque de cohérences logique et chronologique des résumés (ordre des faits cités inapproprié), même si d'autres aspects entrent en ligne de compte : anaphores non résolues, conjonction de coordination indésirables... La section qui suit propose une stratégie pour résoudre ce problème.

3.2 Ordonner les phrases à l'aide des classes lexicales

Nous faisons l'hypothèse que les phrases du résumé peuvent être ordonnées dans le même ordre que les classes dont elles sont issues. Nous faisons également l'hypothèse que l'ordre des phrases dans les documents d'origine reflète un ordre logico-temporel, partagé par l'ensemble des documents à résumer, et qu'il est possible de projeter sur les phrases du résumé. Les méthodes d'ordonnancement que nous proposons se veulent les plus génériques possible, ne faisant appel à aucune technique linguistique dépendante de la langue. Cependant, cette hypothèse contraint les documents sur lesquels la méthode est applicable. Dans l'expérience que nous avons menée, les

2. Le système crée ainsi 8 classes pour extraire 8 phrases d'une longueur moyenne supposée de 22 mots.

documents en entrée sont tous des dépêches de presse, majoritairement organisées autour de la même structuration de l'information, en pyramide inversée.

Les méthodes que nous proposons sont fortement semblables à celle décrite dans (Regina Barzilay *et al.*, 2002). Cependant, celle-ci compte seulement le nombre de documents dans lesquels une phrase d'une classe B suit une phrase d'une classe A, tandis que nous calculons la proportion de phrases de B situées après des phrases de A, transposables sous formes de probabilités. De plus, nous proposons une évaluation différente, en comparant nos méthodes avec une *baseline* plus performante.

3.2.1 Génération du graphe d'ordonnancement

Lors d'une première étape, un graphe orienté d'ordonnancement des phrases est établi, dans lequel les nœuds sont les CL. Le poids des arêtes est calculé de deux manières différentes. Dans la première (méthode DU, pour « Distance Unique »), le poids de l'arête d'une CL A à une CL B est égal à :

$$\frac{n_{A>B} - n_{B>A}}{n_{A>B} + n_{B>A}} \tag{1}$$

où $n_{X>Y}$ est le nombre couples de phrases (p_X, p_Y) tels que :
– p_X appartient à la classe thématique X ;
– p_Y appartient à la classe thématique Y ;
– p_X et p_Y sont extraites du même document ;
– p_X est située après p_Y.
Dans la seconde (méthode DR pour, « Distance Relative »), le poids de l'arête d'une CL A à une CL B est égal à :

$$\frac{\sum_{(p_A,p_B)\in C_p(A,B)} \delta pos(p_A, p_B)}{\sum_{(p_A,p_B)\in C_p(A,B)} |\delta pos(p_A, p_B)|} \tag{2}$$

où :
– $C_p(X, Y)$ est l'ensemble des couples de phrases telles qu'elles appartiennent au même document, la première appartient à la classe X et la seconde à la classe Y.
– $\delta pos(p_X, p_Y)$ est le nombre de phrases séparant p_X de p_Y si p_X est situé après p_Y, et l'opposé du nombre de phrases séparant p_X de p_Y si p_X précède p_Y.

A la différence de DR, DU ne prend pas en compte la distance entre les phrases d'un même document mais uniquement leur position relative. L'idée derrière ces calculs est d'obtenir non seulement l'ordre présumé entre deux groupes de phrases, mais également la probabilité qu'il soit correct.

3.2.2 Parcours du graphe et ordonnancement

Ordonner les nœuds d'un tel graphe est un problème NP-complet. Nous avons donc utilisé une heuristique pour résoudre ce problème : le parcours du graphe débute à partir du nœud qui minimise la somme des valeurs de ses arêtes sortantes, c'est-à-dire le nœud dont les phrases maximisent la probabilité de précéder celles des autres nœuds. Les phrases de ce nœud précèdent donc majoritairement les phrases des autres nœuds. Tant qu'il existe un nœud non visité, dont le lien du nœud courant vers celui-ci est strictement négatif, le parcours se poursuit depuis le nœud qui minimise ce lien (donc celui dont le poids est le plus proche de -1). S'il n'existe pas de tel nœud, l'algorithme parcourt le nœud non visité qui minimise la somme des valeurs de ses arêtes sortantes. L'algorithme se poursuit tant que tous les nœuds du graphe n'ont pas été visités. Les phrases du résumé sont ordonnées dans l'ordre de parcours du graphe.

Au cours des premières expériences que nous avons menées, nous avons constaté que la première phrase choisie par un tel parcours de graphe est souvent inappropriée. Dans un résumé de dépêches, la première phrase doit décrire l'événement clé afin de faciliter la lecture du reste du document. Nous avons donc implémenté une seconde méthode de parcours, qui débute par la phrase dont le score calculé par CBSEAS est le plus important. Cette phrase devrait contenir les mots-clés les plus discriminantants du thème traité, et nous faisons l'hypothèse qu'elle est la plus pertinente comme accroche du résumé.

4 Évaluation

4.1 Baselines

Nous avons implémenté deux *baselines*. La première ordonne les phrases selon le score que CBSEAS leur a attribué. Ainsi, les phrases éventuellement « hors sujet », c'est-à-dire ne correspondant pas à une composante informationnelle importante des documents d'origine, sont repoussées à la fin du résumé et n'en perturbent pas la lecture. Les phrases contenant le vocabulaire le plus fréquent apparaîtront en premier ; dans le cadre de résumés de dépêches, cela peut être bénéfique, car les premières phrases auront une probabilité élevée d'introduire tous les acteurs des événements traités. La deuxième *baseline* consiste à ordonner les phrases selon la date d'émission des documents dont elles sont issues. Si plusieurs phrases appartiennent au même document, elles sont ordonnées selon leur position dans celui-ci.

4.2 Protocole d'évaluation

Nous avons évalué trois systèmes :
– un qui utilise la méthode de génération de graphe DU ;
– un qui utilise la méthode de génération de graphe DR ;
– un dernier qui utilise DU et le parcours de graphe débutant par la phrase la mieux classée par CBSEAS (DU+score).
Ces systèmes ont été comparés aux deux *baselines* présentées en Section 4.1.

L'évaluation a été réalisée sur le corpus RPM2, composé de 200 dépêches de presse, et divisé en 20 thèmes de 10 documents. Chacun des thèmes a été utilisé afin de générer un résumé par CBSEAS, ordonné selon les méthodes DU, DR, DU+score, et les deux *baselines*. Les résumés ont été évalués par quatre évaluateurs différents, en aveugle, selon deux axes : leur cohérence – sur une échelle de 0 à 3, la qualité de l'enchaînement des phrases – et la pertinence de leur première phrase – pertinente (1) ou non (0) [3].

	Score	Date+pos	DU	DR	DU+score
Pourc. accord	35.8%	46.7%	50.8%	44.2%	38.3%
Ecart type moy.	0.74	0.58	0.52	0.69	0.7
Coef. var. moy.	0.68	1.24	0.79	1.08	0.64
Annotateur 1	1.55	0.35	0.85	0.95	1.4
Annotateur 2	1.6	0.75	1.0	0.9	1.5
Annotateur 3	1.35	0.7	0.85	0.95	1.1
Annotateur 4	1.35	0.58	0.75	0.9	0.75
Moyenne	**1.46**	**0.58**	**0.86**	**0.93**	**1.19**

TABLE 1 – Scores de cohérence

	Score	Date+pos	DU	DR	DU+score
Pourc. accord	66.7%	73.3%	70.8%	66.7%	66.7%
Annotateur 1	0.6	0.1	0.45	0.55	0.6
Annotateur 2	0.65	0.25	0.45	0.45	0.65
Annotateur 3	0.75	0.45	0.6	0.7	0.75
Annotateur 4	0.4	0.1	0.35	0.4	0.4
Moyenne	**0.6**	**0.23**	**0.46**	**0.53**	**0.6**

TABLE 2 – Évaluation de la pertinence de la première phrase

Pour effectuer cette annotation, les évaluateurs avaient également pour consigne de ne pas tenir compte de la rhétorique, c'est-à-dire des marqueurs de discours qui assurent le lien logique entre phrases (typiquement des adverbes tels que néanmoins, en outre, cependant...). Ils ne devaient pas non plus pénaliser la redondance inter-phrastique. En revanche, la présence d'expressions non référentielles (anaphores, périphrases...) ininterprétables était à pénaliser.

3. Nous remercions Aleksandra Guerraz et Philippe Fabien pour leur participation à l'évaluation de notre travail.

	score	Date+pos	DU	DR
175 mots	1.35	0.7	0.85	0.95
100 mots	**1.6**	**1**	**1.4**	**1.65**

TABLE 3 – Comparaison des évaluations avec 175 et 100 mots.

4.3 Résultats

Le Tableau 1 présente les résultats de l'évaluation de la cohérence globale des résumés. Les annotateurs ont considéré qu'ordonner les phrases selon leur score produisait les résumés les plus cohérents. La deuxième méthode est DU+score, et la moins bonne la méthode Date+pos. Il est intéressant de constater que le réordonnancement selon le score produit également les accords inter-annotateurs les plus importants. Si le pourcentage d'accord pour cette méthode est faible (le nombre de notations identiques), les notes des annotateurs sont parmi les moins divergentes, avec un coefficient de variation moyen – l'écart type moyen rapporté à la moyenne – de 0.68. La méthode d'ordonnancement la moins probante est la *baseline* 2.

Les résultats témoignent de l'importance du choix de la première phrase. En effet, la méthode DU+score obtient des résultats supérieurs à ceux de la méthode DU seule. Seule changeait dans cette méthode le choix de la première phrase, qui a paru plus judicieux aux annotateurs (*cf* Tableau 2). Les résultats de la méthode DU+score n'égalent toutefois pas ceux de la *baseline* 1, quoique deux annotateurs sur quatre aient jugé les deux méthodes sensiblement de même valeur.

4.4 Conclusion

Les résultats de cette évaluation sont décevants, puisqu'aucun des systèmes présentés n'est meilleur que la *baseline* 1. De plus, les notes sont assez faibles, puisqu'aucun système de réordonnancement n'obtient de note supérieure à la moyenne. Cependant, les résultats obtenus sont à mettre en rapport avec le nombre élevé de mots par résumé utilisé lors de cette expérience – 175. Ce nombre élevé de mots, associé à un nombre de CL également important – 8 – a pour effet d'augmenter le rappel des informations extraites dans les résumés, tout en en diminuant la précision. Ainsi, certaines phrases sont hors-sujet et perturbent la lisibilité du résumé. La Figure 1 présente un même résumé ordonné selon les méthodes *baseline* 1, DU, DU+score et *Baseline* 2. On constate bien le phénomène d'interposition de phrases qui perturbent le discours (ici, la phrase « La tension monte encore d'un cran entre Rachida Dati et les magistrats. »).

Une évaluation préliminaire, réalisée avec une configuration différente, avait livré des conclusions différentes. Les résumés étaient limités à 100 mots, et générés à partir de 5 CL. Un seul annotateur avait participé à cette étude. Le Tableau 3 montre que la méthode DU avait été jugée la plus efficace, légèrement devant la *baseline* 1. Pour comparaison, nous avons présenté dans ce tableau l'évaluation de cet annotateur sur les résumés de 175 mots. On constate que la tâche se complexifie rapidement avec l'accroissement du nombre de phrases ; la seule méthode qui reste stable est la *baseline* 1. Afin d'éviter, comme c'est le cas dans beaucoup des résumés de 175 mots générés pour notre expérience, que des phrases trop marginales ne perturbent la lecture du résumé, on peut imaginer prendre en compte le score des phrases du résumé lors du calcul du graphe, de manière à affaiblir les arêtes qui relient les nœuds dont la phrase représentative est mal notée par CBSEAS. Ainsi, la probabilité que de telles phrases soient reléguées en fin de résumé sera plus importante.

Bien que les résultats de cette étude soient décevants, nous avons identifié le principal problème relatif aux méthodes proposées, et suggéré une manière d'y remédier, qu'il conviendra que nous évaluions.

Références

BOSSARD A. & DE NEEF E. (2011). Etude du regroupement automatique de phrases sur un système de résumé automatique. In *Actes de CORIA 2011*, Avignon, France.

BOSSARD A. & POIBEAU T. (2009). Description of the LIPN Systems at TAC 2009. In *TAC 2009, Summarization Track*, Gaithersburg, USA.

DAS D. & MARTINS, ANDRÉ F. T. (2007). A survey on automatic text summarization.

Méthode : *baseline* 1, score : 2.25
Irrités par la politique de la garde des Sceaux, les deux principaux syndicats ont saisi le Conseil supérieur de la magistrature (CSM) pour « l'alerter » sur « les attaques » de la ministre « contre l'indépendance de l'autorité judiciaire ».
L'organe suprême de la magistrature était saisi par les deux principaux syndicats de magistrats, l'Union syndicale des magistrats (USM, majoritaire) et le Syndicat de la magistrature (SM, gauche).
A la suite du suicide d'un détenu mineur à la maison d'arrêt de Metz, le garde des Sceaux a diligenté une enquête administrative qui selon les magistrats "vise à trouver un bouc émissaire".
Les magistrats et les organisations du monde pénitentiaire ont manifesté ce jeudi pour protester contre la politique de la Garde des Sceaux Rachida Dati. Comme d'autres syndicats et les pétitionnaires d'Agen, l'USM fait le lien entre l'histoire de Bernard Blais et la récente convocation au ministère du vice-procureur de Nancy.
La tension monte encore d'un cran entre Rachida Dati et les magistrats.

Méthode : DU, score : 0.75
Irrités par la politique de la garde des Sceaux, les deux principaux syndicats ont saisi le Conseil supérieur de la magistrature (CSM) pour « l'alerter » sur « les attaques » de la ministre « contre l'indépendance de l'autorité judiciaire ».
Comme d'autres syndicats et les pétitionnaires d'Agen, l'USM fait le lien entre l'histoire de Bernard Blais et la récente convocation au ministère du vice-procureur de Nancy.
L'organe suprême de la magistrature était saisi par les deux principaux syndicats de magistrats, l'Union syndicale des magistrats (USM, majoritaire) et le Syndicat de la magistrature (SM, gauche).
A la suite du suicide d'un détenu mineur à la maison d'arrêt de Metz, le garde des Sceaux a diligenté une enquête administrative qui selon les magistrats "vise à trouver un bouc émissaire".
La tension monte encore d'un cran entre Rachida Dati et les magistrats.
Les magistrats et les organisations du monde pénitentiaire ont manifesté ce jeudi pour protester contre la politique de la Garde des Sceaux Rachida Dati.

Méthode : DU+score, score : 1.75
Irrités par la politique de la garde des Sceaux, les deux principaux syndicats ont saisi le Conseil supérieur de la magistrature (CSM) pour « l'alerter » sur « les attaques » de la ministre « contre l'indépendance de l'autorité judiciaire ».
Comme d'autres syndicats et les pétitionnaires d'Agen, l'USM fait le lien entre l'histoire de Bernard Blais et la récente convocation au ministère du vice-procureur de Nancy.
L'organe suprême de la magistrature était saisi par les deux principaux syndicats de magistrats, l'Union syndicale des magistrats (USM, majoritaire) et le Syndicat de la magistrature (SM, gauche).
A la suite du suicide d'un détenu mineur à la maison d'arrêt de Metz, le garde des Sceaux a diligenté une enquête administrative qui selon les magistrats "vise à trouver un bouc émissaire".
La tension monte encore d'un cran entre Rachida Dati et les magistrats.
Les magistrats et les organisations du monde pénitentiaire ont manifesté ce jeudi pour protester contre la politique de la Garde des Sceaux Rachida Dati.

Méthode : Date+pos, score : 0
L'organe suprême de la magistrature était saisi par les deux principaux syndicats de magistrats, l'Union syndicale des magistrats (USM, majoritaire) et le Syndicat de la magistrature (SM, gauche).
La tension monte encore d'un cran entre Rachida Dati et les magistrats.
Irrités par la politique de la garde des Sceaux, les deux principaux syndicats ont saisi le Conseil supérieur de la magistrature (CSM) pour « l'alerter » sur « les attaques » de la ministre « contre l'indépendance de l'autorité judiciaire ».
Comme d'autres syndicats et les pétitionnaires d'Agen, l'USM fait le lien entre l'histoire de Bernard Blais et la récente convocation au ministère du vice-procureur de Nancy.
Les magistrats et les organisations du monde pénitentiaire ont manifesté ce jeudi pour protester contre la politique de la Garde des Sceaux Rachida Dati. A la suite du suicide d'un détenu mineur à la maison d'arrêt de Metz, le garde des Sceaux a diligenté une enquête administrative qui selon les magistrats "vise à trouver un bouc émissaire".

FIGURE 1 – Résumés ordonnés selon les méthodes *baseline* 1, DU, DU+score et *Baseline* 2.

DE LOUPY C., GUÉGAN M., AYACHE C., SENG S. & TORRES MORENO J.-M. (2010). A french human reference corpus for multi-document summarization and sentence compression. In *LREC'10*.

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*.

MANI I. (1999). *Advances in Automatic Text Summarization*. Cambridge, MA, USA : MIT Press.

REGINA BARZILAY, NOEMIE ELHADAD & KATHLEEN MCKEOWN (2002). Inferring Strategies for Sentence Ordering in Multidocument News Summarization. *J. Artif. Intell. Res. (JAIR)*, **17**, 35–55.

RUTLEDGE L., BAILEY B., OSSENBRUGGEN J. V., HARDMAN L. & GEURTS J. (2000). Generating presentation constraints from rhetorical structure. In *In Proceedings of the 11th ACM conference on Hypertext and Hypermedia*.

SAGGION H. & GAIZAUSKAS R. (2004). Multi-document summarization by cluster/profile relevance and redundancy removal. In *Proceedings of the Document Understanding Conference 2004* : NIST.

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