Génération de résumés de mise à jour : Utilisation d'un algorithme de classification non supervisée pour détecter la nouveauté dans les articles de presse

Aurélien Bossard

Génération de résumés de mise à jour : Utilisation d'un algorithme de classification non supervisée pour détecter la nouveauté dans les articles de presse

Aurélien Bossard

Laboratoire d'Informatique de Paris-Nord
UMR 7030, CNRS et Université Paris 13)
99, av. J.-B. Clment
93430 Villetaneuse
prenom.nom@lipn.univ-paris13.fr

Résumé. Dans cet article, nous présentons un système de résumé automatique multi-documents, dédié au résumé de mise à jour – ou de nouveauté. Dans une première partie, nous présentons la méthode sur laquelle notre système est fondé, CBSEAS, et son adaptation à la tâche de résumé de mise à jour. Générer des résumés de mise à jour est une tâche plus compliquée que de générer des résumés « standard », et nécessite une évaluation spécifique. Nous décrivons ensuite la tâche « Résumé de mise à jour » de TAC 2009, à laquelle nous avons participé afin d'évaluer notre système. Cette campagne d'évaluation internationale nous a permis de confronter notre système à d'autres systèmes de résumé automatique. Finalement, nous présentons et discutons les résultats intéressants obtenus par notre système.

1 Introduction

La recherche en résumé automatique, supportée par des campagnes d'évaluations et une communauté de chercheurs importantes, a connu ces dernières années des progrès rapides tant du point de vue des méthodes employées que des résultats qualitatifs. En effet, ce domaine répond à des besoins forts en recherche et extraction d'information, dûs notamment à l'augmentation des données électroniques consultables. Le domaine du résumé automatique s'étend, et s'intéresse désormais à différents médias comme supports de différents types de résumé. Comme champs d'application, on peut citer les fils d'e-mails, les blogs, les articles scientifiques ou encore les articles de presse ; et comme types de résumé les synthèses d'opinion, les résumés différentiels et les résumés de mise à jour. C'est précisément ce dernier type de résumé auquel nous nous intéressons dans cet article.

Un résumé de mise à jour sur un sujet donné est fondé sur l'hypothèse que le lecteur du résumé a déjà lu certains documents concernant ce sujet. Le résumé de mise à jour doit synthétiser les informations dont l'utilisateur n'a pas encore pris connaissance, donc les informations nouvelles qui apparaissent dans de nouveaux documents.

Dans cet article, nous présentons nos recherches en résumé automatique de mise à jour : nous avons développé un système de résumé automatique « standard », et l'avons adapté au résumé de mise à jour. Ce système est fondé sur une classification automatique de phrases à résumer, qui permet d'augmenter la diversité informationnelle des synthèses générées.

La tâche de génération de résumés de mise à jour a été proposée durant les campagnes d'évaluation DUC[1] 2007, et TAC[2] 2008 et 2009. Nous avons participé à la tâche « *Update Task* » de la campagne TAC 2009. Cette tâche permet d'évaluer deux types différents de résumé : les résumés « standard », guidés par une requête, et les résumés de mise à jour, également guidés par une requête.

Dans cet article, nous nous fondons sur un système de résumé automatique multi-documents, CBSEAS (Bossard, 2009), qui diffère des autres systèmes de résumé par l'identification de sous-thèmes et l'utilisation de la centralité locale à un sous-thème, en plus de la centralité globale, afin de produire de meilleurs résumés. Le regroupement de phrases véhiculant les mêmes informations, et donc leur classification en sous-thèmes, est crucial pour le résumé automatique multi-documents : savoir quelles phrases sont différentes les unes des autres, mais également parmi celles qui sont similaires, détecter la phrase centrale, peut aider à produire des résumer avec une diversité informationnelle importante. Au-delà de la question de la génération du résumé automatique, la question principale à laquelle nous tentons de répondre ici est : comment distinguer les informations nouvelles de celles déjà lues ?

Cet article montre comment les différents aspects de ces deux problématiques sont gérés. Nous présentons également l'évaluation de notre système sur la tâche « Résumé de mise à jour » de TAC 2008 et 2009.

2 État de l'art

Dans cette section, nous présentons un aperçu des méthodes existantes pour le résumé automatique et la gestion de la mise à jour. Ces domaines ont été largement étudiés ; nous limitons donc cet état de l'art aux travaux principaux et à ceux qui ont le plus inspiré notre approche.

2.1 Les approches de résumé automatique multi-documents

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 des documents à résumer, et en

[1] Document Understanding Conference : http://www-nlpir.nist.gov/projects/duc/index.html
[2] Text Analysis Conference : http://www.nist.gov/tac

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 (1969) a défini des indices textuels qui peuvent être utilisés afin de déterminer l'importance d'une phrase. Il a notamment proposé une liste de mots-clés tels que « *hardly* », « *In conclusion* ». Les indices contiennent également la position des phrases et le nombre de mots qui co-occurrent dans le titre du document. Ces indices sont encore utilisés de nos jours dans la majorité des systèmes de résumés automatiques, comme dans celui de (Kupiec et al., 1995), qui les combine à un algorithme d'apprentissage. Cependant, ceux-ci sont limités puisqu'ils ne prennent pas en compte le contenu global du document.

D'autres systèmes se concentrent sur les fréquences des termes en corpus. Luhn (1958) a ouvert la voie aux systèmes statistiques de résumé par extraction. Il a proposé de construire une liste des termes importants des documents, en se fondant sur leur fréquence. Sont sélectionnés seulement ceux dont la fréquence appartient à un intervalle prédéfini. Plus une phrase présente de mots appartenant à cette liste, plus elle est pertinente. Radev et al. (2002) ont profité des avancées dans le domaine des statistiques textuelles en intégrant le *tf.idf* à la méthode de Luhn. La liste des termes importants, que Radev appelle « centroïde », est composée des *n* termes avec le plus grand *tf.idf*. Les phrases sont classées selon leur similarité au centroïde.

Les méthodes statistiques sont efficaces pour sélectionner les phrases qui reflètent le contenu global des documents à résumer. Une telle phrase est qualifiée de « centrale ». Cependant, ces méthodes ne sont pas conçues de manière à générer des résumés qui reflètent la diversité informationnelle des documents d'origine. La diversité informationnelle est aussi importante que la centralité lorsqu'on évalue la qualité d'un résumé. En effet, un résumé doit contenir toutes les informations importantes.

La méthode MMR – *Maximum Margin Relevance* – Carbonell et Goldstein (1998) cherche à résoudre le problème de la diversité. Les phrases qui maximisent la fonction de score présentée en Équation 1 sont sélectionnées incrémentalement. La fonction de score MMR prend en compte la diversité en soustrayant au score la centralité la similarité maximale de la phrase évaluée avec les phrases déjà sélectionnées. Cette méthode est utilisée largement et a été adaptée à différentes tâches de résumé automatique (Goldstein et al., 2000; Chowdary et Kumar, 2009; Ribeiro et de Matos, 2007; Wang et al., 2009).

$$MMR = \underset{P_i \in D \setminus S}{\operatorname{argmax}} \left[\lambda sim_1(P_i, Q) - (1 - \lambda) \underset{P_j \in S}{\operatorname{argmax}}\ sim_2(P_i, P_j) \right] \tag{1}$$

où Q est la requête utilisateur, D l'ensemble des phrases, S l'ensemble des phrases sélectionnées pour le résumé, et λ le facteur de nouveauté.

Dans le cas particulier du résumé multi-documents, la redondance est un bon indice de l'importance d'un élément d'information. MMR prend en compte la redondance, mais seulement dans le but d'éliminer les phrases redondantes, et non comme un critère d'extraction. Radev (Erkan et Radev, 2004) s'est appuyé sur les avancées récentes dans le domaine des réseaux sociaux afin d'utiliser la redondance comme le principal critère pour juger la pertinence d'une phrase. Il construit un graphe des documents à résumer, dans lequel les nœuds qui ont

le plus grand *prestige* sont ceux qui sont fortement liés à d'autres nœuds ayant eux-mêmes un *prestige* important.

Toutes les méthodes que nous avons présentées dans cet état de l'art considèrent le contenu global des documents à résumer pour évaluer la centralité des phrases. Cependant, nous considérons les documents non comme un tout, mais comme différents groupes de phrases formant des sous-thèmes. Dans chacun de ces sous-thèmes, des phrases centrales émergent, qui sont celles que nous voulons extraire.

2.2 Les approches de résumé de mise à jour

La tâche « Résumé de mise à jour » de DUC 2007 et TAC 2008 a révélé que générer un résumé de mise à jour est une tâche bien plus complexe que de générer des résumés « standard » Dang et Owczarzak (2008). Cette tâche pose en effet, au-delà du problème de la génération d'un résumé automatique, celui de la détection de la nouveauté. Nous présentons ici différentes stratégies visant à gérer le résumé de mise à jour.

Certaines auteurs, comme Galanis et Malakasiotis (2008), retirent des documents de mise à jour toutes les phrases dont la similarité à une phrase des documents initiaux est supérieure à un seuil défini empiriquement. D'autres préfèrent supprimer les phrases qui maximisent la similarité au jeu de documents initial jusqu'à ce que la similarité globale entre ce dernier et le jeu de documents de mise à jour soit en dessous d'un seuil prédéfini (He et al., 2008).

La méthode présentée par Boudin et Torres-Moreno (2008) sélectionne les phrases pour le résumé de mise à jour en utilisant la méthode MMR, décrite en Section 2.1. Le poids de la similarité aux phrases déjà sélectionnées est augmenté afin de réduire le risque d'extraire des phrases qui ne véhiculent pas d'information nouvelle.

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 dans les documents postérieurs :

$$fn(w) = \frac{|nd_t|}{|pd_t| + |D|} \tag{2}$$

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

Le facteur de nouveauté est utilisé pour mesurer la nouveauté d'une phrase. Cette méthode a prouvé son efficacité, tant sur les évaluations de TAC 2008 que sur celles de TAC 2009. Cependant, nous voulons évaluer une nouvelle méthode fondée sur la similarité entre phrases, qui ne nécessite pas, contrairement aux premières approches présentées, de fixer *a priori* un seuil de similarité.

3 CBSEAS, une approche générique pour le résumé automatique multi-documents

Nous voulons gérer spécifiquement l'aspect multi-documents en considérant la redondance comme le problème principal du résumé multi-documents. En effet, nous considérons les documents à résumer comme constitués de groupes de phrases qui véhiculent la même information. Dans chacun de ces groupes, une phrase peut être considérée centrale. Extraire une phrase dans chaque groupe de phrases peut mener à réduire le risque de voir apparaître de la redondance dans les résumés générés. De plus, extraire la phrase centrale permet de prendre en compte la centralité locale de chaque sous-thème. Enfin, cette modélisation nous autorise à prendre en compte un critère supplémentaire pour extraire les phrases : la centralité d'une phrase vis-à-vis du sous-thème dans lequel elle a été classée.

Notre système implémente cette méthode. La première étape consiste à regrouper les phrases similaires, puis d'extraire une phrase par classe.

3.1 Regroupement de phrases

Cette section décrit la première partie de notre système : le regroupement de phrases. Nous voulons un algorithme de classification flexible, dans lequel nous pouvons aisément adapter le critère de regroupement. *Fast global k-means* apparaît approprié à cet effet : cet algorithme prend en entrée une matrice de similarité ou de distance. Le modèle créé après avoir regroupé les phrases peut être utilisé non seulement afin d'extraire les phrases, mais également à des fins d'ordonnancement des phrases. Ceci sera l'objet de futures publications.

3.1.1 Pré-traitements

Les documents en entrée subissent des pré-traitements avant d'être traités par CBSEAS. Nous présentons ici les différents pré-traitements réalisés.

Étiquetage morpho-syntaxique Les documents sont analysés morpho-syntaxiquement : l'étiquetage morpho-syntaxique est assuré par tree-tagger[3] (Schmid, 1994). Cela permet de prendre en compte les différents types morpho-syntaxiques pendant le calcul de la similarité entre phrases.

Segmentation en phrases Certains auteurs choisissent de travailler sur des petites structures textuelles plutôt que sur des phrases complètes. Ils travaillent donc à l'extraction de groupes de mots syntaxiquement liés, et divisent les phrases en propositions (Marcu, 1998). Extraire des propositions plutôt que des phrases pose le problème de l'identification de telles propositions – bien que l'analyse syntaxique automatique ait récemment connu d'importants progrès – et de leur indépendance. D'autres auteurs extraient des paragraphes entiers dans le but d'augmenter la cohérence linguistique des résumés. Cependant, cela augmente le risque d'extraire des phrases non pertinentes.

Nous avons donc fait le choix d'extraire des phrases entières afin d'éviter de générer des résumés agrammaticaux et d'extraire des phrases non pertinentes.

[3] page web de tree-tagger : http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/

$$sim(s_1, s_2) = \frac{\sum_{mt} weight(mt) \times fsim(s_1, s_2)}{\frac{fsim(s_1, s_2) + gsim(s_1, s_2)}{\sum_{mt} weight(mt)}} \quad (3)$$

$$fsim(s_1, s_2) = \sum_{n_1 \in s_1} \sum_{n_2 \in s_2} tsim(n_1, n_2) \times \frac{tfidf(n_1) + tfidf(n_2)}{2} \quad (4)$$

$$gsim(s_1, s_2) = card\left((n_1 \in s_1, n_2 \in s_2) \mid tsim(n_1, n_2) < \delta\right) \quad (5)$$

where mt are the morphological types, s_1 and s_2 the sentences, tsim the similarity between two terms using WordNet and the JCn similarity measure Jiang et Conrath (1997) and δ a similarity threshold.

Étiquetage en entités nommées L'étiquetage en entités nommées permet de raffiner le calcul de similarité entre phrases. En effet, la reconnaissance de telles entités autorise la prise en compte de groupes lexicaux complexes tels que « le Président George W. Bush » comme une seule et même entité lexicale. Un tel groupe doit en effet être identifié comme une seule entité nommée. Dans le calcul de similarité entre phrases, il sera donc considéré comme un terme unique, et non comme quatre termes distincts. Les entités nommées sont étiquetées par le système ANNIE (Cunningham et al., 2002), développé pour l'architecture GATE.

3.1.2 Calcul de similarité entre phrases

Nous faisons l'hypothèse que la similarité entre phrases doit prendre en compte le type de documents que CBSEAS doit résumer, ainsi que le type de résumé demandé par l'utilisateur. Par exemple, les caractéristiques qui déterminent si deux phrases sont similaires diffèrent selon que l'on cherche à générer un résumé d'opinions ou un résumé d'analyse boursière. Dans le premier cas, les adjectifs, adverbes et verbes de sentiments sont discriminants ; dans le second cas, les catégories discriminantes seront les devises, montants, et noms de compagnie, soit majoritairement des entités nommées.

Nous voulons prendre en compte ce fait en utilisant une mesure de similarité paramétrable, qui peut être aisément adaptée aux différentes tâches auxquelles un système de résumé automatique peut être confronté. Nous voulons également prendre en compte les relations linguistiques entre termes – *e.g.* la synonymie, l'hyperonymie. Nous utilisons pour cela la mesure de similarité JCn (Jiang et Conrath, 1997) qui est fondée sur la distance entre *synsets* dans la taxonomie de WordNet (Fellbaum, 1998). Les Équations 3, 4, 5 présentent cette mesure.

3.1.3 Algorithme de classification

Une fois la matrice de similarité calculée, CBSEAS regroupe automatiquement les phrases similaires. Cette étape est réalisée en utilisant *fast global k-means* (Likas et al., 2001), une variante incrémentale de l'algorithme des k-moyennes (MacQueen, 1967). *Fast global k-means* résoud le problème du choix des *k* centres de classe initiaux posé par l'algorithme des k-moyennes. L'incrémentalité de *fast global k-means* le rend également intéressant dans le but de générer des résumés de mise à jour. Bien que des méthodes de classification ascendante

- p1 : Chess legend **Bobby Fischer**, who faces prison if **Bobby** returns to the **United States**, can only avoid **deportation** from **Japan** if **Iceland** upgrades its granting of **residency** to full citizenship, a **Japanese** lawmaker said Wednesday.
- p2 : **Fischer**, 62, has been detained in **Japan** awaiting **deportation** to the **United States**, where **Bobby** is wanted for violating economic sanctions against the former Yugoslavia by playing a highly publicized chess match there in 1992.
- p3 : **Fischer**, 62, has been detained in **Japan** awaiting **deportation** to the **United States**, where **Bobby** is wanted for violating economic sanctions against the former Yugoslavia by playing a highly publicized chess match there in 1992.
- p4 : But during Thursday's meeting the committee said it was awaiting more documents from **Japan** to be sure **Bobby** would be let go if given citizenship in **Iceland**.
- p5 : Palsson said **Fischer**, 62, would be informed of the vote Tuesday morning in **Japan**, where he has been detained awaiting deportation to the **United States** on a warrant for violating economic sanctions against the former Yugoslavia by playing a chess match there in 1992.
- p6 : **Iceland** recently issued **Fischer** a special passport and **residence** permit to help resolve the standoff over **Bobby**'s status.
- p7 : **Iceland** had previously sent **Fischer** a passport which showed **residency** status but not full nationality, a document not deemed sufficient to save him from **deportation** to the **United States**.
- p8 : **Iceland** had already agreed to issue a **resident**'s permit and special passport to **Fischer**, but that had not proved enough for the **Japanese** authorities.

Source : données de TAC 2009, Tâche « Résumé de mise à jour »

FIG. 1 – *Exemple d'une classe générée par CBSEAS*

hiérarchique aient été testées (et n'ont pas égalé les performances de *fast global k-means*), nous avons jusqu'à présent fourni une part plus importante de travail sur la question du critère de regroupement, donc de la similarité entre phrases (Bossard et Guimier De Neef, 2011). Celle-ci a en effet des conséquences non seulement sur la qualité du regroupement, mais également sur la sélection des phrases (*cf* Section 3.2.1).

Fast global k-means crée d'abord une classe qui contient tous les éléments à classifier. A chaque itération, l'algorithme ajoute une nouvelle classe dont le centre est l'élément le moins représentatif de sa classe. Chaque élément est alors placé dans la classe dont il est le plus proche du centre, et le centre de chaque classe est recalculé. L'algorithme s'arrête lorsque le nombre de classes demandé par l'utilisateur est atteint.

La Figure 1 présente une classe générée par CBSEAS. Les mots partagés par au moins la moitié des phrases de cette classe sont en gras afin d'identifier les raisons de ce regroupement.

3.2 Sélection de phrases

Après avoir regroupé les phrases, CBSEAS extrait une phrase par classe. Rappelons que les phrases ainsi extraites doivent minimiser la redondance, et ainsi produire un résumé avec une bonne diversité informationnelle. Le critère d'extraction est aussi important que le regroupement de phrases. En effet, la méthode utilisée pour déterminer les phrases à extraire influence la centralité du résumé. Nous présentons ici les trois critères utilisés par CBSEAS pour extraire les phrases : centralités globale et locale, et longueur des phrases. Le score final d'une phrase est la somme pondérée de ces trois scores. Les poids sont fixés à l'aide d'un algorithme génétique, détaillé dans (Bossard et Rodrigues, 2011).

3.2.1 Centralité locale

La centralité locale est la pertinence d'une phrase vis-à-vis du contenu de sa classe (ou sous-thème). Nous voulons que les phrases extraites reflètent au mieux les informations de

Liste des informations atomiques (AI) véhiculées par les phrases en Fig. 1 :

Information	**Poids**
Fischer faces deportation	5
Ficher chess player	4
Iceland issued Fischer a resident permit	4
Fischer violated economic sanctions against Yugoslavia	3
Iceland issued Fischer a passport	3
Fischer is 62	3
Fischer has been detained in Japan	3
Fischer faces prison	1
Fischer could avoid deportation	1
Iceland special passport can not avoid him deportation	1
Iceland wants to be sure Fischer would be let go if given citizenship	1

p2, p3 et p5 sont les phrases véhiculant le plus d'informations centrales :

Phrase	p1	p2	p3	p4	p5	p6	p7	p8
Somme des poids des AI	15	21	21	1	21	7	8	7

Similarités entre les phrases :

	p1	p2	p3	p4	p5	p6	p7	p8	somme
p1	1.0	.171	.171	.156	.150	.1	.133	.133	2.014
p2	.171	1.0	1.0	.091	.821	.031	.094	.064	3.272
p3	.171	1.0	1.0	.091	.821	.031	.094	.064	3.272
p4	.156	.091	.091	1.0	.083	.075	.069	.077	1.642
p5	.15	.821	.821	.083	1.0	.027	.108	.055	3.065
p6	.1	.031	.031	.075	.027	1.0	.167	.315	1.746
p7	.133	.094	.094	.069	.108	.167	1.0	.115	1.78
p8	.133	.064	.053	.077	.055	.315	.115	1.0	1.812

Scores de centralité locale tels que définis en Section 3.2.1 :

Phrase	p1	p2	p3	p4	p5	p6	p7	p8
Score	.171	1.0	1.0	.091	.821	.031	.094	.064

FIG. 2 – *Illustration du concept de centralité locale*

leur classe. L'idée derrière la centralité locale est la suivante : les phrases d'une classe C expriment un ensemble d'informations atomiques AI. La phrase la plus centrale de C est celle qui contient les informations les plus importantes de AI. Nous travaillons selon l'hypothèse que la redondance d'une information est corrélée à son importance. La phrase qui maximise la somme des similarités aux autres phrases, P_{max}, est la phrase la plus centrale. Elle reçoit un score de centralité égal à 1. Les autres phrases reçoivent un score égale à leur similarité à P_{max}. La Figure 2 illustre ce calcul.

3.2.2 Centralité globale

Le problème majeur d'une extraction de phrases fondée exclusivement sur la mesure de la centralité locale est la non prise en compte du contenu global des documents ou d'une éventuelle requête utilisateur. Pour générer des résumés précis qui répondent au besoin informationnel exprimé par un utilisateur, ou qui sont centrés sur les thèmes principaux des documents à résumer, nous devons ajouter la mesure de la centralité globale à celle de la centralité locale. Pour cela, nous identifions deux cas :

– l'utilisateur a une requête, le résumé doit donc être en rapport avec celle-ci ;
– l'utilisateur n'a pas de requête, le résumé doit donc être pertinent vis-à-vis du contenu global des documents.

Dans le premier cas, nous utilisons la similarité à la requête comme score de centralité globale. Celui-ci est calculé de la même manière que la similarité entre phrases, présentée en Section 3.1.2. Dans le second cas, nous utilisons le score centroïde, tel que défini dans (Radev et al., 2002).

3.2.3 Longueur des phrases

La longueur des résumés est souvent limitée à un certain nombre de mots. Pour cette raison, nous avons choisi de donner un score à chaque phrase, en fonction de leur longueur, dans le but de pénaliser les phrases trop petites ou trop longues. La fonction de ce score est présentée en Équation 6.

$$score_{longueur} = \frac{1}{e^{(|longueur(phrase) - longueur_{demandee}|)}} \tag{6}$$

4 Générer des résumés de mise à jour

Avec le développement des sites de news en ligne, la détection et le résumé de la nouveauté est devenue une problématique de recherche importante. En effet, les utilisateurs qui suivent un sujet ne veulent pas avoir à lire tout nouvel article publié, mais uniquement les informations dont ils n'ont pas déjà pris connaissance. Le résumé de mise à jour répond à des besoins importants pour l'accès au contenu. De plus , les systèmes de résumé automatique « standard » génèrent des synthèses qui véhiculent un contenu informationnel qualitativement correct. La recherche peut donc se concentrer sur des tâches plus complexes, telles que celles récemment proposées par les campagnes d'évaluation DUC et TAC : résumé d'opinions, résumé de mise à jour (ou de nouveauté), ou résumé thématique. Dans cette section, nous présentons notre méthode pour gérer le résumé de mise à jour.

4.1 Intuitions

CBSEAS – Clustering-Based Sentence Extractor for Automatic Summarization – 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 véhiculent des informations connues ;
– celles qui véhiculent des informations nouvelles.

Nous partons donc de l'hypothèse que les phrases des nouveaux documents qui véhiculent des informations anciennes partagent le même vocabulaire (étendu grâce à la similarité JCn) que les phrases que l'utilisateur a déjà lues.

La principale faiblesse d'une telle méthode est le manque de traitements linguistiques dédiés. Par exemple, la phrase *« After hemming and hawing and bobbing and weaving, the board of directors of Fanni Mae finally jettisoned Franklin D. Raines, the mortgage finance giant's former executive, and Timothy Howard, its former chief financial officer. »* issue du corpus

AQUAINT-2[4] est aisément identifiable comme une phrase porteuse de nouveauté du fait du temps employé et de l'emploi de « *finally* ». Dans le cas spécifique des dépêches de presse, cela signifie que l'information vient tout juste de paraître. Cependant, si l'utilisation d'indices linguistiques peut aider à détecter les phrases porteuses de nouveauté, cela limite également la méthode à un langage et un domaine uniques.

De plus, CBSEAS a démontré son efficacité à regrouper des phrases sémantiquement proches et à différencier les phrases éloignées. En effet, CBSEAS s'est classé à la troisième place tous systèmes confondus pour la gestion de la redondance dans les résumés sur la tâche « Résumé d'opinions » de TAC 2008 Bossard et al. (2008). C'est une raison supplémentaire pour utiliser notre méthode de regroupement afin de détecter les phrases porteuses de nouveauté.

4.2 Algorithme de mise à jour

Avant de tenter d'identifier les phrases porteuses de nouveauté, nous devons modéliser les informations que l'utilisateur a déjà lues. Nous pouvons alors confronter les nouveaux documents à ce modèle afin de déterminer si les phrases de ces documents véhiculent des informations nouvelles. La première étape de notre algorithme consiste donc à classifier les phrases issues des documents déjà lus – que nous appelons D_I – en k_I classes, comme décrit en Section 3.1.3.

Le modèle ainsi calculé – M_I – est alors utilisé pour la seconde étape de notre algorithme, qui consiste à déterminer si une phrase des nouveaux documents – D_U – doit être regroupée avec les phrases de D_I à compléter, ou créer une nouvelle classe qui ne contiendra que des phrases porteuses de nouveauté. *Fast global k-means*, moyennant quelques adaptations, peut être utilisé pour confronter des éléments à un modèle précédemment établi dans le but de déterminer si ces éléments peuvent intégrer ce modèle. Nous décrivons ici la partie de notre algorithme dédiée à la détection de nouveauté.

Premièrement, les similarités entre les phrases de D_U et les centres de classe de M_I ainsi qu'entre toutes les phrases de D_U sont calculées. Alors, les phrases de D_U à compléter sont 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 déplacées vers un autre cluster, ceci afin de préserver le modèle M_I qui encode les anciennes informations. Cela évite également de perturber la portée sémantique des nouvelles classes, qui sont porteuses de nouveauté.
- Les centres de classe de M_I ne sont pas recalculés ; étant donné que la portée sémantique d'une classe dépend directement de son centre, cela évite de modifier la portée sémantique des classes de M_I par ajout de nouveaux éléments issus de D_U.

Le principal défaut de cet algorithme, qui est détaillé dans la Figure 3 est le choix de k_I – le nombre de classes de M_I – et celui de k_U – le nombre de classes de M_U. Nous avons décidé empiriquement de fixer k_U au nombre de phrases désiré pour le résumé de mise à jour, et k_I à $\frac{P_I}{P_U} \times k_U$, où P_I et P_U sont respectivement les phrases de D_I et D_U. Pour notre participation à la tâche « Résumé de mise à jour » de TAC 2009, nous avons utilisé un algorithme génétique entraîné sur les données de TAC 2008 afin de décider des meilleures

[4]Le corpus AQUAINT-2 est un sous-ensemble de la troisième édition du corpus anglais LDC Gigaword composé de nouveaux articles issus de différentes agences de presse.

//Classification pour M_I
pour tous les p de P_I
faire
  $cluster(p) \leftarrow C_1$
fin pour
pour i de 1 à k_I
faire
  pour n de 1 à i
  faire
    $centre(C_n) \leftarrow \underset{p_j \in C_n}{\text{argmax}}(\sum_{pm \in C_n} sim(p_j, p_m))$
  fin pour
  pour tous les p de P_I
  faire
    $cluster(p) \leftarrow \underset{C_m, 1<m<u}{\text{argmax}} (sim(centre(C_m), p)$
  fin pour
  si $i < k_I$ alors
    $cluster(\underset{p \in D_I}{\text{argmin}}(sim(p, centre(cluster(p)))) \leftarrow C_{i+1}$
  fin si
fin pour
//Détection de la nouveauté
pour tous les p de P_U
faire
  $cluster(p) \leftarrow argmax_{C_i, 1<i<k_I}(sim(centre(C_i), p))$
fin pour
pour i de k_I à $k_I + k_U$
faire
  pour n de $k_I + 1$ à i
  faire
    $centre(C_n) \leftarrow \text{argmax}_{p_j \in C_n}(\sum_{pm \in C_n} sim(p_j, p_m))$
  fin pour
  pour tous les p dans P_U
  faire
    $cluster(p) \leftarrow \underset{C_m, 1<m<i}{\text{argmax}} (sim(centre(C_m), p))$
  fin pour
  si $i < k_U$ alors
    $cluster(\text{argmin}_{p_m \in D_I}(sim(p_m, centre(cluster(p)))) \leftarrow C_{i+1}$
  fin si
  fin pour

FIG. 3 – *Algorithme de détection de la mise à jour*

valeurs pour ces variables. Aucune de ces solutions n'est idéale, puisqu'elles nécessitent qu'il existe au moins k_U nouvelles informations véhiculées par au moins k_U phrases différentes. D'autres solutions peuvent être envisagées, comme déterminer si ajouter des classes de mise à jour améliore ou détériore la qualité de la classification, en utilisant un indice de validité (Davies et Bouldin, 1979; Calinski et Harabasz, 1974; Beale, 1969).

Une fois les classes peuplées, le résumé de mise à jour est généré en extrayant une phrase par classe de mise à jour, comme décrit en Section 3.2.

5 Évaluation : Participation à TAC

Nous avons évalué notre travail sur la tâche de « Résumé de mise à jour » de la campagne d'évaluation TAC 2009, organisée par le NIST[5]. Nous présentons en détails la tâche, les différentes méthodes d'évaluation, et les résultats obtenus par notre système de résumé de mise à jour.

5.1 Description détaillée de la tâche

Les participants à la tâche de « Résumé de mise à jour » de la campagne d'évaluation de TAC 2009 devaient produire deux types différents de résumé : les résumés « standard » et les « résumés de mise à jour », tous deux guidés par une requête.

La tâche consiste en 44 sujets qui comportent chacun un titre (court), une requête, ainsi que deux jeux de documents : les documents initiaux et les documents de mise à jour. Les systèmes doivent générer deux résumés pour chacun des sujets : un résumé « standard » qui synthétise l'information des documents de mise à jour. Ce dernier résumé doit être produit en tenant compte du fait que son lecteur a déjà pris connaissance du contenu des documents initiaux. Les résumés sont limités à une longueur de 100 mots, quelle que soit la taille des documents d'origine.

Chaque jeu de documents comprend dix documents extraits du corpus ACQUAINT-2 (*cf* Section 4.1). Ces documents sont des dépêches de presse en anglais issues de différentes sources : AFP, NYT, APW, LTW, et Xinhua.

Les requêtes sont en langue anglaise et peuvent être complexes. Si la requête associée au Sujet D0848, présenté dans la Figure 4 est simple, ce n'est pas le cas de toutes, comme celle du sujet D0902 : « *Describe the debate over use of emergency contraceptives, also called the morning-after pill, and whether or not it should be available without a prescription* ».

5.2 Méthodes d'évaluation

Le NIST a utilisé trois méthodes différentes afin d'évaluer les résumés des participants. Les résumés sont évalués par le *package* d'évaluation ROUGE[6] Lin (2004). Les métriques ROUGE sont fondées sur les co-occurrences de n-grammes entre des résumés de référence et les résumés à évaluer. Leur principal avantage réside dans leur fonctionnement entièrement automatique. ROUGE peut donc être utilisé pour des expérimentations en dehors des campagnes d'évaluation. Cependant, l'évaluation de résumés ne peut pas être limitée à des comparaisons de séquences de n-grammes. NIST a donc choisi d'utiliser des méthodes d'évaluation plus précises mais non automatiques.

La seconde méthode utilisée par le NIST est la méthode Pyramide, décrite en détails dans (Nenkova et al., 2007). Les auteurs définissent la notion de SCU – *Summarization Content Unit* – une information qui apparaît dans les résumés. La méthode Pyramide consiste à extraire une liste de SCUs depuis les résumés de référence. Ces SCUs sont alors classés selon leur nombre d'occurrences. Selon les auteurs, ce classement peut être vu comme une pyramide où les informations les plus importantes sont au sommet et les moins importantes à la base. Les

[5]NIST : National Institute of Standards and Technology

[6]ROUGE : Recall-Oriented Understudy for Gisting Evaluation

Sujet D0848 : Airbus A380		Describe developments in the production and launch of the Airbus A380
Documents initiaux		
16/01/2005	AFP	The Airbus A380 : from drawing board to runway-ready in a decade
16/01/2005	AFP	A380 'superjumbo' will be profitable from 2008 : Airbus chief
16/01/2005	APW	Airbus prepares to unveil 1380 « superjumbo », world's biggest passenger jet
17/01/2005	LTW	Can Airports Accomodate the Giant Airbus A380 ?
19/01/2005	AFP	After fanfare, Airbus A380 now must prove it can fly
25/01/2005	AFP	Airbus mulls boosting A380 production capacity
10/04/2005	AFP	While US government moans, airports ready for Airbus giant
27/04/2005	AFP	Paris airport neighbors complain about noise from giant Airbus A380
27/04/2005	NYT	Giant Airbus 380 makes maiden flight
04/05/2005	AFP	Airbus A380 takes off on second test flight
Documents de mise à jour		
01/06/2005	AFP	Airbus announces delay in delivering new superjumbo A380
03/06/2005	AFP	German wing of Airbus denies superjumbo A380 parts delay
05/10/2005	AFP	US aviation officials to study A380 turbulence
15/10/2005	AFP	Airbus says it cannot meet demand for A380 superjumbo
18/10/2005	APW	Second Airbus A380 makes maiden flight
13/11/2005	APW	Airbus executive says company will pay millions in compensation for late A380deliveries
17/02/2006	APW	Airbus sees no delay to A380 after wing ruptured during test
22/02/2006	AFP	Airbus confident of A380 certification
26/03/2006	APW	33 people injured in evacuation frill for A380 super-jumbo
29/03/2006	APW	Airbus A380 superjumbo passes emergency evacuation test

FIG. 4 – *Exemple d'un sujet issu de la tâche « Résumé de mise à jour » de TAC 2009.*

résumés sont finalement évalués en extrayant les SCUs qu'ils contiennent et en les comparant à la pyramide.

L'évaluation Pyramide prend en compte la qualité linguistique des résumés : si une phrase est agrammaticale, elle ne véhicule aucun ou peu de SCUs. Cependant, cette méthode d'évaluation ne prend pas en compte la cohérence globale du résumé. C'est la raison pour laquelle le NIST a introduit des mesures d'évaluation entièrement manuelles. Elles sont décrites dans (Dang et Owczarzak, 2009). Ces mesures manuelles évaluent la « performance générale » (*overall responsiveness sic.*) et la lisibilité. La performance générale mesure la qualité linguistique, et à quel point un résumé répond aux besoins informationnels. Le score de lisibilité reflète la grammaticalité, la non-redondance, la clarté référentielle, le *focus*, la structure et la cohérence. La performance générale et la lisibilité ont été évaluées sur une grille à cinq points :

– 5 : très bon
– 4 : bon
– 3 : acceptable
– 2 : mauvais
– 1 : très mauvais.

Il aurait été intéressant de disposer d'une évaluation des différents critères sur lesquels le score de lisibilité est fondé, comme c'était le cas pour la tâche « Résumé d'opinion » de TAC 2008. Cela nous aurait permis de mieux analyser les résultats de notre système.

5.3 Baselines

Le NIST a fourni trois *baselines* pour la tâche « Résumé de mise à jour » de TAC 2009. La première (notée *Baseline* 1 dans les résultats) est le résultat de l'extraction des premières phrases dans le document le plus récent, jusqu'à ce que la limite de 100 mots soit atteinte. Cette *baseline* fournit une limite inférieure de la qualité que doit atteindre un système de résumé automatique plus probante que la sélection aléatoire de phrases.

La seconde *baseline* (Baseline 2) est générée en ré-ordonnant aléatoirement les phrases d'un résumé de référence. Cela donne un aperçu de l'impact d'un mauvais ordonnancement des phrases sur la qualité linguistique et la performance générale des résumés.

La troisième *baseline* (Baseline 3) est constituée de phrases complètes extraites manuellement. La méthode d'extraction est détaillée dans (Genest et al., 2009). L'idée sous-jacente à cette *baseline* est d'évaluer la limite suéprieure de ce que peut générer un système de résumé automatique, tant du point de vue du contenu que de la qualité linguistique.

5.4 Résultats et discussion

Dans cette section, nous présentons les résultats de notre système, les comparons à ceux des autres participants et les discutons.

La Figure 6 présente les résultats des évaluations Pyramide et Performance générale pour tous les participants. Notre système se classe parmi le premier tiers des participants pour les résumés initiaux, et dans les dix meilleurs systèmes pour les résumés de mise à jour. Le score Performance générale n'est pas aussi bon. Cela est dû à la mauvaise qualité linguistique des résumés générés par notre système. En effet, CBSEAS n'applique aucun des post-traitements communément appliqués, qui pourraient améliorer la cohérence des résumés. La Figure 8 présente les deux résumés générés par CBSEAS pour le Sujet D0911. La dernière phrase est coupée. Cela est dû à la limite de 100 mots imposée par la tâche. CBSEAS ne retire pas automatiquement la phrase qui dépasse cette limite. Cela a également un effet négatif sur le score de Qualité linguistique.

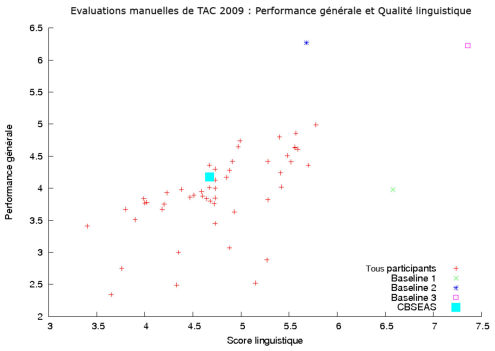

FIG. 5 – *Performance générale et score linguistique des systèmes de TAC 2009*

La Figure 7 présente les différents scores obtenus par CBSEAS, et sa position vis-à-vis des autres systèmes. La qualité linguistique apparaît comme le véritable point faible de notre

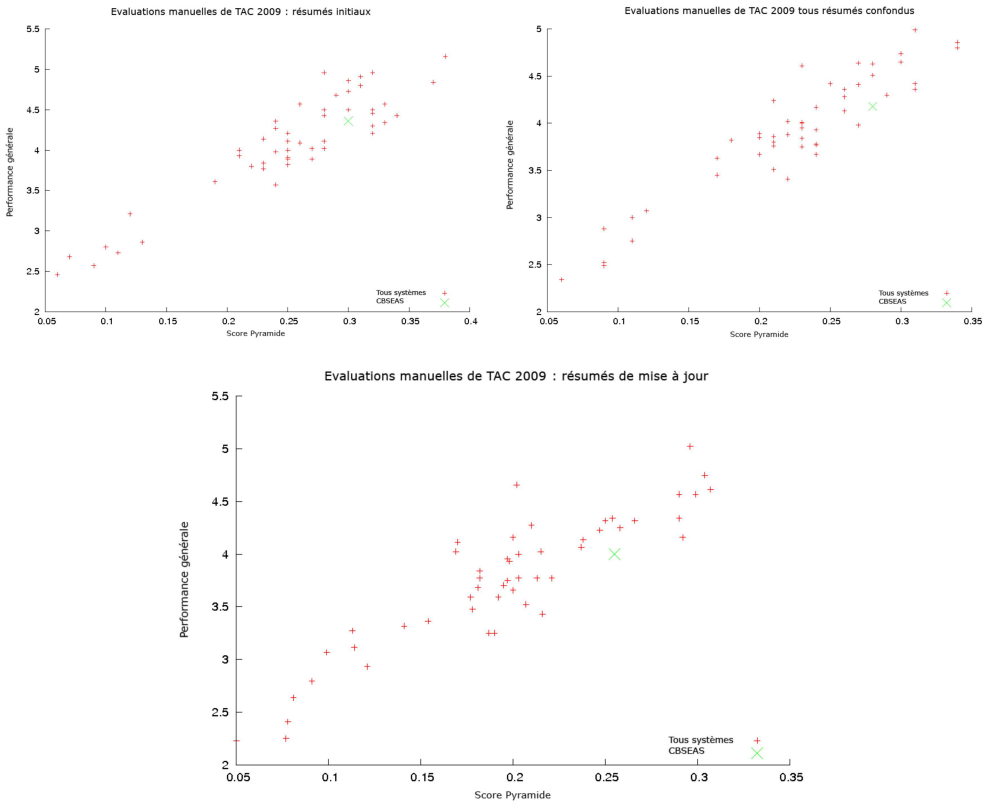

FIG. 6 – *Résultats de la tâche « Résumé et Mise à jour » de TAC 2009 : scores Pyramide et Performance générale*

système. Cependant, cette vue témoigne de l'efficacité de notre stratégie de gestion de la mise à jour. Cinq systèmes sur les quinze qui devancent CBSEAS pour le score Pyramide des résumés initiaux sont moins bien classés que notre système pour le score Pyramide des résumés de mise à jour. Les résumés de mise à jour obtiennent en moyenne des scores Pyramide 21.5% inférieurs à ceux des résumés initiaux. Notre système perd « seulement » 13.5% en score Pyramide – qui évalue le contenu informationnel. D'une manière générale, il a été noté que le résumé de mise à jour est une tâche délicate, et l'on peut constater que les systèmes produisent des résumés initiaux de meilleure qualité que les résumés de mise à jour.

L'évaluation proposée n'est pas complète : les résumés initiaux et de mise à jour sont évalués indépendamment les uns des autres. L'évaluation des résumés de mise à jour aurait pu être poussée plus avant, en évaluant notamment la présence de SCUs dans les résumés de mise à jour qui sont présents dans les documents initiaux. La redondance entre les résumés de mise à jour et le contenu déjà connu de l'utilisation n'est en effet pas évaluée.

Le score linguistique de la *Baseline 2*, présenté en Figure 5, est intéressant : égal à 5.68, il est dépassé par deux systèmes automatiques. Les meilleurs systèmes atteignent des scores

Scores moyens des résumés initiaux et de mise à jour

	ROUGE-2	ROUGE-SU4	Pyr.	Ling.	Perf. gén.
Class. de CBSEAS	9/53	10/53	11/53	31/53	18/53
Score de CBSEAS	0.0919	0.1305	0.28	4.67	4.18
Meilleur score	0.0273	0.0583	0.06	3.40	2.34
Moins bon score	0.1127	0.1452	0.34	5.78	4.99
Score moyen	0.0786	0.1168	0.226	4.751	3.922

Résumés initiaux

	ROUGE-2	ROUGE-SU4	Pyr.	Ling.	Perf. gén.
Class. de CBSEAS	8/53	8/53	15/53	35/53	19/53
Score de CBSEAS	0.1027	0.1338	0.3	4.91	4.3
Moins bon score	0.0282	0.0591	0.06	3.43	2.46
Meilleur score	0.1216	0.1510	0.38	5.93	5.16
Score moyen	0.0853	0.1214	0.252	4.762	4.075

Résumés de mise à jour

	ROUGE-2	ROUGE-SU4	Pyr.	Ling.	Perf. gén.
Class. de CBSEAS	8/53	15/53	10/53	24/53	22/53
Score de CBSEAS	0.0811	0.1223	0.26	4.75	3.98
Moins bon score	0.0264	0.0576	0.05	3.36	2.23
Meilleur score	0.1039	0.1395	0.31	5.89	5.02
Score moyen	0.0719	0.1122	0.198	4.742	3.769

FIG. 7 – *Résultats numériques détaillés de la tâche « Résumé et mise à jour » de TAC 2009*

équivalents à la *Baseline* 3 – constituée de phrases extraites manuellement – pour ce qui est de sélectionner les informations les plus importantes (score Pyramide). Cependant, ces systèmes sont toujours loin derrière cette *baseline* (*cf* Figure 5) pour la qualité linguistique et la performance générale. Cela prouve l'impact du réordonnancement de phrases sur la qualité linguistique, mais également sur la satisfaction d'un utilisateur vis-à-vis d'un résumé, exprimée par le score de performance générale.

La campagne d'évaluation TAC 2009 a montré que notre système est compétitif pour générer des résumés au contenu informationnel important, mais ne produit pas des résumés d'une qualité linguistique à la hauteur de leur contenu. CBSEAS arrive en effet à 83.3 % du score Pyramide de la *Baseline 3* – qui fournit le maximum de ce que pourrait réaliser un système de résumé automatique. La gestion de la mise à jour est très satisfaisante, puisque CBSEAS se classe encore mieux sur cette tâche que sur la tâche de résumé « standard ».

6 Conclusion

Dans cet article, nous avons présenté CBSEAS, un système générique de résumé automatique multi-documents, et un nouvel algorithme dédié à la gestion des résumés de mise à jour – ou de nouveauté. Notre système a obtenu des résultats compétitifs pendant la campagne d'évaluation TAC 2009. Les résultats comparés des résumés « standard » et des « résumés de mise à jour » montrent que notre stratégie de gestion de la mise à jour est efficace. Cependant, elle

pourrait être améliorée en filtrant en amont de notre méthode, les phrases des documents de mise à jour en utilisant une méthode telle que celle décrite dans Varma et al. (2009), fondée sur le facteur de nouveauté des mots. Les résultats mettent aussi en avant la qualité de la méthode de sélection des phrases par CBSEAS. Cependant, notre système manque de post-traitements, ce qui influe négativement sur la satisfaction générale des utilisateurs. Si le réordonnancement de phrases semble être particulièrement influent sur la compréhension du résumé par un utilisateur, l'impact d'autres post-traitements tels que la compression de phrases ou la résolution d'anaphores devraient être évalué dans des travaux à venir.

D0911 Bobby Fischer : résumé initial
Describe efforts to secure asylum in Iceland for chess legend Bobby Fischer.
Chess legend Bobby Fischer was on Monday granted citizenship by the parliament of Iceland, a move which could allow him to avoid deportation from Japan to the United States where he is wanted for violating sanctions against the former Yugoslavia. Chess legent Bobby Fischer, who faces prison if he returns to the United States, can only avoid deportation from Japan if Iceland upgrades its granting of residency to full citizenship. Iceland's parliament voted Monday to grant citizenship to fugitive U.S. chess star Bobby Fischer. Lawmakers in Iceland are likely to grant citizenship to mercurial chess genius Bobby Fischer, a
Score Pyramide : 0.622 Score linguistique : 6

D0911 Bobby Fischer : résumé de mise à jour
Describe efforts to secure asylum in Iceland for chess legend Bobby Fischer.
Iceland said Wednesday it hoped to give detained chess legend Bobby Fischer a passport before the weekend after granting him citizenship in a move that could allow him to avoid a US prison term. An Icelandic supporter of Bobby Fischer said Tuesday he had paid a registration fee that would allow the American chess legend to settle in Iceland. Chess legend Bobby Fischer could leave his Japanese detention cell by the weekend, his supporters said Tuesday, a day after Iceland's parliament voted to grant him citizenship. Japan said Tuesday it may let detained chess legend Bobby Fischer leave for Iceland,
Score Pyramide : 0.345 Score linguistique : 5

FIG. 8 – *Exemple d'un couple de résumés générés par CBSEAS.*

Références

Beale, E. M. L. (1969). Euclidean cluster analysis. *Bulletin of the International Statistical Institute 43*, 92–94.

Bossard, A. (2009). CBSEAS, a new approach to automatic summarization. In *Proceedings of the SIGIR 2009 Conference - Doctoral Consortium*, Boston, USA.

Bossard, A., M. Généreux, et T. Poibeau (2008). Description of the lipn systems at tac2008 : Summarizing information and opinions. In *Notebook papers and results of TAC 2008*, Gaithersburg, Maryland, USA.

Bossard, A. et E. Guimier De Neef (2011). étude de l'impact du regroupement automatique de phrases sur un système de résumé multi-documents. Technical report.

Bossard, A. et C. Rodrigues (2011). Combining a multi-document update summarization system – cbseas – with a genetic algorithm. Smart Innovation, Systems and Technologies. Springer.

Boudin, F. et E.-B. M. Torres-Moreno, Juan-Manuel (2008). A scalable MMR approach to sentence scoring for multi-document update summarization. In *Proceedings of the 2008 COLING Conference*, Manchester, UK, pp. 21–24.

Calinski, R. B. et J. Harabasz (1974). A dendrite method for cluster analysis. *Communications in Statistics 3*, 1–27.

Carbonell, J. et J. Goldstein (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*, New York, NY, USA, pp. 335–336. ACM.

Chowdary, C. R. et P. S. Kumar (2009). Esum : An efficient system for query-specific multi-document summarization. In *Proceedings of the 31th European Conference on IR Research on Advances in Information Retrieval*, ECIR '09, Berlin, Heidelberg, pp. 724–728. Springer-Verlag.

Cunningham, H., D. Maynard, K. Bontcheva, et V. Tablan (2002). GATE : A framework and graphical development environment for robust NLP tools and applications. In *Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics*, Philadelphia, PA, USA.

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

Dang, H. T. et K. Owczarzak (2009). Overview of the TAC 2009 update summarization task. In *Notebook papers and results of TAC 2009*, Gaithersburg, Maryland, USA.

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

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

Fellbaum, C. (1998). *WordNet : An Electronic Lexical Database*.

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

Genest, P.-É., G. Lapalme, et M. Yousfi-Monod (2009). Hextac : the creation of a manual extractive run. In *Notebook papers and results of TAC 2009*, Gaithersburg, Maryland, USA.

Goldstein, J., V. Mittal, J. Carbonell, et M. Kantrowitz (2000). Multi-document summarization by sentence extraction. In *NAACL-ANLP 2000 Workshop on Automatic summarization - Volume 4*, Morristown, NJ, USA, pp. 40–48. Association for Computational Linguistics.

He, T., J. Chen, Z. Gui, et F. Li (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.

Jiang, J. J. et D. W. Conrath (1997). Semantic similarity based on corpus statistics and lexical taxonomy. In *International Conference Research on Computational Linguistics (ROCLING X)*.

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

Likas, A., N. Vlassis, , et J. Verbeek (2001). The global k-means clustering algorithm. *Pattern Recognition 36*, 451–461.

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.

Luhn, H. (1958). The automatic creation of literature abstracts. *IBM Journal 2*(2), 159–165.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In L. M. Le Cam et J. Neyman (Eds.), *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, Volume 1, Statistics. University of California Press.

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

Nenkova, A., R. J. Passonneau, et K. McKeown (2007). The pyramid method : Incorporating human content selection variation in summarization evaluation. *TSLP 4*(2).

Radev, D., A. Winkel, et M. Topper. (2002). Multi document centroid-based text summarization. In *Proceedings of the ACL 2002 Demo Session*, Philadelphia, PA, USA.

Ribeiro, R. et D. M. de Matos (2007). Extractive summarization of broadcast news : comparing strategies for european portuguese. In *Proceedings of the 10th international conference on Text, speech and dialogue*, TSD'07, Berlin, Heidelberg, pp. 115–122. Springer-Verlag.

Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the International Conference on New Methods in Language Processing*, Manchester, UK.

Varma, V., P. Bysani, K. R. V. Bharat, S. Kovelamudi, S. GSK, K. Kumar, et N. Maganti (2009). Iit hyderabad at tac 2009. In *Notebook papers and results of TAC 2009*, Gaithersburg, Maryland, USA.

Wang, B., B. Liu, C. Sun, X. Wang, et B. Li (2009). Adaptive maximum marginal relevance based multi-email summarization. In *Proceedings of the International Conference on Artificial Intelligence and Computational Intelligence*, AICI '09, Berlin, Heidelberg, pp. 417–424. Springer-Verlag.

Générer des résumés de mise à jour

Summary

In this article, we present a summarization system dedicated to update summarization. We first present the method on which this system is based, CBSEAS, and its adaptation to the update summarization task. Generating update summaries is a far more complicated task than generating "standard" summaries, and needs a specific evaluation. We describe TAC 2009 "Update Task", which we used in order to evaluate our system. This international evaluation campaign allowed us to confront CBSEAS to others automatic summarization systems. Finally, we show and discuss the interesting results obtained by our system.