Filtrage de textes dans le but de produire un résumé de documents multiples

Fatma Kallel Jaoua*, — Lamia Hadrich Belguith*,***— Maher Jaoua*,*** — Abdelmajid Ben Hamadou *,******

* *Laboratoire MIRACL,*
Institut Supérieur d'informatique et de Multimédia de Sfax, 3000 Sfax, Tunisie
** *Fatma_fseg@yahoo.fr* , *** *{l.belguith, Maher.Jaoua}@fsegs.rnu.tn*
**** *Abdelmajid.benhamadou@isimsf.rnu.tn*

RÉSUMÉ. *Dans le cadre de la conférence d'évaluation DUC, nous avons développé un système de résumé automatique de documents multiples qui se base sur l'extraction des phrases clés. La méthode proposée utilise un algorithme génétique qui permet de combiner les phrases des documents sources pour former les extraits, qui seront croisés et mutés pour générer de nouveaux extraits. L'examen des résultats obtenus dans les deux sessions DUC'04 et DUC'07 a montré un écart significatif au niveau des performances du système développé. En effet, un phénomène de dérive génétique est observé lorsqu'on traite, en entrée de notre système, des textes de grande taille. Afin de remédier à cette dérive, nous proposons d'intégrer un module supplémentaire de filtrage qui a pour objectif la réduction du nombre des phrases des textes sources en entrée. Ce filtrage est effectué sur la base de la notion de dominance entre phrases qui permet d'éliminer un grand nombre de phrases du pool initial.*

ABSTRACT. *In the context of DUC Conference (Document Understanding Conference) , we have developed an automatic summarization system of multiple documents which is based on the extraction of the key sentences. The proposed method uses a genetic algorithm which combines the sentences of the source documents in order to produce extracts. These extracts will be crossed and mutated in order to generate new extracts. The examination of the results obtained in the two sessions DUC' 04 and DUC' 07 showed a significant variation of the system performance. Indeed, a phenomenon of genetic drift is observed when the system processes big size texts (as an input). In order to solve this problem, we propose to integrate an additional module of sentence filtering to reduce the number of sentences in the input. This filtering is based on the concept of predominance between sentences which allows to eliminate a great number of sentences from the initial pool.*

MOTS-CLÉS : *Résumé automatique de documents multiples, filtrage de texte, dominance entre phrases, algorithme génétique, mémoire génétique.*

KEYWORDS : *Automatic summarization of multiple documents, text filtering, sentence dominance, genetic algorithm, genetic memory .*

1. Introduction

Les travaux de recherche développés dans le cadre du résumé automatique ont été, initialement, orientés vers l'automatisation de la condensation des documents simples. Néanmoins, et suite à l'explosion documentaire et à l'abondance des collections de documents traitant des thèmes similaires, on a ressenti la nécessité de mettre au point de nouveaux outils automatiques permettant la réduction et le filtrage des informations utiles à partir de documents multiples. Cette automatisation vise, essentiellement, à faciliter la recherche et l'extraction de l'information textuelle pertinente en vue d'assurer l'assimilation des documents obtenus.

C'est dans ce contexte, que nous avons proposé une méthode de résumé automatique de documents multiples décrivant un même thème. Cette méthode se base sur l'extrait en tant qu'unité minimale d'extraction. La méthode proposée opère par constitution d'un ensemble d'extraits qui sont, ensuite, évalués et classés en vue de déterminer le meilleur en tenant compte de certains critères statistiques et linguistiques. La mise en œuvre de cette méthode illustrée par le système ExtraNews, a été marquée par l'utilisation d'un algorithme génétique qui simule le processus de génération et de classement d'extraits.

La version du système ExtraNews traitant la langue anglaise, a été évaluée dans les conférences d'évaluation DUC'04[1] (Jaoua et *al*., 2004) et DUC'07. L'examen des résultats de ces évaluations, a montré une divergence au niveau des résultats obtenus. En effet, les premiers rangs occupés par notre système lors de l'évaluation du contenu au niveau de la conférence DUC'04 sont en discordance avec les résultats obtenus dans la conférence DUC'07. Malgré la différence des tâches et les mesures d'évaluations employées dans les deux sessions d'évaluation, la divergence des résultats met en cause le choix des étapes de mise en œuvre de notre système. En effet, une dérive génétique est observée dans le cadre de la conférence DUC'07, ce qui explique les résultats obtenus par notre système. Cette dérive se manifeste par une altération de la fonction "objectif" utilisée pour le classement des extraits des populations générées par l'application de l'algorithme génétique.

Afin de remédier à cette dérive, nous proposons dans cet article l'intégration d'un module supplémentaire de filtrage qui permet de réduire l'espace de recherche exploré par l'algorithme génétique. Nous proposons entre autres, l'utilisation d'une mémoire génétique permettant de mémoriser les meilleures solutions (extraits) récoltées au fil des générations produites. Ce qui permet d'éviter la dérive et de pouvoir ainsi explorer un plus grand espace de recherche.

Cet article s'articule autour de cinq sections. Cette partie introductive est suivie par la deuxième section qui présente un tour d'horizon des travaux réalisés dans le

1. DUC : Document Understanding Conference : est une compagne d'évaluation internationale des systèmes de résumé automatique de documents en langue anglaise. La compagne DUC sera remplacé à partir de l'année 2008 par la compagne TAC : Text Analysis Conference. http://duc.nist.gov

domaine du résumé automatique de documents multiples. La troisième section détaille les fondements de base de la méthode proposée ainsi que l'architecture initiale de notre système. La quatrième section est dédiée à l'étude qui porte sur l'évaluation menée lors de la conférence DUC'04 et DUC'07. La cinquième section décrit les améliorations apportées à notre système ainsi que les résultats de l'évaluation menée après l'intégration de ces améliorations.

2. Etat de l'art

La plupart des travaux réalisés dans le domaine du résumé automatique de documents multiples s'articulent autour du processus de regroupement et de classement des unités textuelles similaires en vue de dégager les unités les plus importantes. L'importance des unités textuelles est, généralement, représentée par un poids ou une probabilité qui leur est assignée en fonction de leur richesse en mots clés, de leurs positions dans le document, etc. (Minel, 2002).

Dans la section suivante, nous présentons un bref aperçu des principales méthodes proposées dans le cadre du domaine du résumé automatique de documents multiples.

2.1. Méthodes statistiques

La plupart des méthodes statistiques se basent sur l'attribution d'une pondération aux phrases pour identifier celles qui participent à l'élaboration du résumé. Dans ce cadre, Carbonell et Goldstein ont proposé une nouvelle métrique appelée MMR (Relevance Maximale Marginale) dans le but de réduire la redondance tout en maximisant la diversité des passages sélectionnés (Carbonell et *al*., 1998). La valeur MMR d'une phrase dépend de sa similarité par rapport à une requête (ou une question) de l'utilisateur et de sa dissimilitude avec les phrases qui ont été déjà sélectionnées dans le résumé.

Pour calculer le poids des phrases des documents sources, Lin et *al* ont proposé une méthode qui combine la valeur MMR avec des techniques de résumé de documents simples (Lin et *al*., 2002). Ils ont combiné plusieurs heuristiques à savoir la position de la phrase, la fréquence des termes, la signature des concepts et le groupement de termes.

La formule de calcul de poids des phrases peut intégrer d'autres heuristiques liées à l'importance d'un document par rapport à la collection des documents sources. Dans ce cadre, Mori et *al*., ont exploité une nouvelle métrique appelée le ratio de gain informationnel (IGR) (Mori et *al*., 2004). Pour déterminer ce ratio, Mori et *al*. ont établit une hiérarchie de classes entre les documents en se basant sur un calcul de similarité. Ils déterminent pour chaque mot un poids qui dépend de la consistance de sa distribution dans la hiérarchie des classes. L'importance d'une

phrase dépend principalement des scores IGR des mots qu'elle contient ainsi que du calcul de la relevance marginale maximale (MMR).

2.2. Méthodes à base de connaissances linguistiques

Afin de repérer les unités textuelles similaires, certaines méthodes se sont basées sur des connaissances linguistiques pour éviter la sélection des unités redondantes dans le résumé à générer. Ces connaissances ont pour objectif de calculer les coréférences, de déterminer les entités nommées, d'appréhender les formes de cohésion et ce afin de distinguer les phrases pertinentes.

Saggion et *al*., ont proposé une méthode de résumé de documents traitant de la biographie humaine et qui se base sur le classement des phrases en utilisant des métriques de coréférence et d'informations lexicales (Saggion et *al*., 2004). Une mesure de similarité est déterminée pour chaque couple de phrases jugées pertinentes afin de réduire la redondance.

Fuentes s'est basée sur des connaissances linguistiques pour extraire les formes de cohésions reliant les phrases des textes sources (Fuentes, 2003). Elle utilise à cet effet les chaînes lexicales, les chaînes de co-références et les chaînes des entités nommées. Chaque phrase est pondérée en fonction des chaînes qui la traversent tout en identifiant les phrases similaires.

La méthode proposée par McKeown s'appuie sur une analyse linguistique et statistique en vue d'extraire des heuristiques statistiques et lexicales de l'ensemble des documents en entrée (McKeown, 1999). Sur la base de ces caractéristiques, cette méthode détermine les phrases du résumé qui véhiculent les thèmes similaires. Ces phrases sont, ensuite, ordonnées en fonction de la date de l'émission de leurs articles sources et combinées moyennant des règles grammaticales. Cette méthode a été implémentée dans le système SimFinder présenté dans la conférence DUC'01 (McKeown, 2001). SimFinder se distingue par l'emploi de certaines heuristiques de mesure et de regroupements flexibles qui participent à la détermination des groupes d'unités semblables du texte (phrases ou paragraphes). Il permet, grâce à un module de sélection et de génération, de réduire chacun de ces groupes en une seule phrase. Dans un travail récent, des techniques de fusion des phrases jugées pertinentes ont été intégrées dans ce système pour améliorer la qualité du résumé généré (Barzilay *et al*., 2005).

2.3. Méthodes par instanciation de *templates*

Ces méthodes se distinguent par l'utilisation de modèles prédéfinis de résumé. Ainsi, il s'agit de remplir les slots du modèle (*template*) choisi. Nous citons dans ce cadre, les travaux de Radev et *al* qui ont proposé une méthode de génération automatique de résumé basée sur l'instanciation de *templates* à partir des

informations similaires extraites des articles sources (Radev et al., 1998). Ils utilisent des règles discursives pour détecter les cas de similarité, d'évolution, de contradiction et de généralisation entre les concepts qui substituent les slots du *template* prédéfini. Cette méthode se propose de générer le résumé final en se basant sur les *slots* du *template* ainsi que sur les relations entre les concepts. White et al utilisent le même principe mais, en employant des heuristiques permettant de regrouper les phrases dans une première étape, puis de calculer leurs poids afin de déterminer la correspondance des phrases possédant les poids les plus importants avec les slots des *templates* (White 2001).

2.4. Méthodes par compréhension

Les méthodes par compréhension consistent à développer une représentation interne des textes sources en vue de détecter les composantes importantes qui sont ensuite réduites puis reformulées pour former le résumé final. Dans ce cadre, Mani et al., ont cherché à exprimer, dans l'extrait, les différences qui peuvent s'établir entre les documents en entrée (Mani et al., 1997). Afin de déterminer ces informations, ils représentent chaque document par un graphe conceptuel où les termes clés représentent les nœuds de ce graphe, et les arcs représentent la relation entre ces termes. Ils utilisent, ensuite, un algorithme de propagation afin de déterminer les nœuds liés sémantiquement au sujet traité par les documents sources. Les graphes activés de deux documents sont combinés afin de trouver un graphe correspondant aux similarités et aux différences entre chaque paire de documents. Il est à noter que cette technique exige des connaissances préalables du domaine à traiter pour extraire correctement l'information.

3. Méthode proposée

L'examen des travaux réalisés met en relief deux approches de résumé automatique. La première approche se base sur l'extraction des phrases porteuses d'idées clés. Alors que la seconde se propose de générer des résumés en se basant sur une analyse plus ou moins profonde des textes sources (Minel, 2002). Une approche mixte peut être utilisée pour générer automatiquement le résumé : dans une première étape, il s'agit de sélectionner un extrait constitué à partir des phrases clés des documents sources. La deuxième étape consiste à générer un résumé en appliquant des mécanismes de révision sur l'extrait produit par la première étape. Dans cet article, nous nous intéressons particulièrement à détailler les améliorations que nous avons apportées à la première étape qui se propose de produire un extrait renfermant les idées clés véhiculées par les documents sources. Actuellement, nous avons entamé les travaux de passage vers la deuxième étape en intégrant des mécanismes de révision de l'extrait.

Afin d'appréhender le problème d'extraction des phrases pertinentes dans un document, nous avons exploré une nouvelle unité d'extraction qui opère à un niveau plus large que la phrase. Ce choix est guidé par le fait qu'un niveau englobant la phrase permet de mieux contrôler les problèmes résultant de la sélection des phrases indépendamment les unes des autres (Jaoua et *al*. , 2003). Le niveau d'extraction choisi est l'extrait qui est formé à partir de phrases des documents sources. Ainsi le processus d'extraction est vu comme étant un problème d'optimisation où il s'agit d'effectuer une comparaison entre plusieurs extraits en vue de sélectionner le meilleur. Il s'agit, donc, de choisir à partir des textes sources un sous ensemble de phrases répondant à un certain nombre de critères liés à la qualité et à la quantité des informations véhiculées.

Toutefois, la détermination de l'ensemble des partitions d'un document est un problème NP (non polynomial) qui ne peut pas être résolu en un temps raisonnable (Brucker et *al*., 1978). Afin de résoudre ces problèmes, les méthodes d'optimisation opèrent par évaluation de solutions intermédiaires en vue de converger vers une solution "optimale". L'application de ces méthodes pour le problème d'extraction suppose que toutes les solutions intermédiaires représentent des extraits générés en une première étape, puis évalués en fonction des critères utilisés.

Dans le contexte d'extraction des phrases clés, la méthode d'optimisation proposée par Jaoua et *al*. (Jaoua et *al*., 2003) préconise l'utilisation d'un algorithme génétique pour explorer l'ensemble des partitions formées par la concaténation des phrases des textes sources. Le choix des algorithmes génétiques est essentiellement motivé par la grandeur de l'espace de recherche. En effet, l'utilisation des algorithmes génétiques dans le cadre de l'extraction offre la possibilité de travailler sur plusieurs solutions en même temps, ce qui permet d'explorer un grand espace de recherche. Cela n'est pas possible avec les méthodes exactes.

3.1. Mise en oeuvre du système ExtraNews

3.1.1. Description des modules du système ExtraNews

La mise en œuvre du système Extranews s'articule principalement autour du module de génération et de classement. Ce module utilise un algorithme génétique qui permet de générer « aléatoirement » en une première étape, un ensemble d'extraits qui sont ensuite évalués et classés. Le classement utilise les résultats de deux autres modules à savoir le module statistique et le module linguistique et qui ont pour tâche la détermination des mots clés dans les documents sources. L'architecture de notre système a distingué les modules suivants :

– Le module statistique : ce module permet de calculer la fréquence des mots présents dans les différents documents sources. Pour ce faire, nous avons défini trois listes de mots clés : la liste de mots fréquents (après avoir éliminé les mots outils) et la liste des mots issus des titres des documents sources et la liste issue de la question posée par l'utilisateur. Ces trois listes forment la liste des mots clés qui va servir, par

la suite, à la détermination de la couverture et de la pondération des phrases des documents en entrée.

– Le module linguistique : le module linguistique permet d'enrichir la liste des mots clés issus des documents ou de la requête utilisateur par des mots synonymes. Pour cela, il utilise le dictionnaire WordNet qui permet d'invoquer des fonctions prédéfinies en vue d'obtenir la liste des synonymes d'un mot (Fellbaum, 1998). L'enrichissement de la liste des mots clés permet aussi de corriger leurs fréquences. Ainsi, si deux mots clés synonymes apparaissent dans les textes sources, leurs fréquences s'additionnent.

– Le module de génération aléatoire et de classement : ce module s'intéresse à la génération et au classement des extraits moyennant une évaluation multicritère. Il utilise un algorithme génétique (voir section suivante) qui permet d'explorer et de comparer une multitude de solutions. L'algorithme se base sur le principe de génération aléatoire d'une population de génomes qui sera classée en fonction d'une valeur d'adaptation. Dans notre cas, le génome constitue l'extrait alors que la phrase représente un gène de ce génome. Les meilleurs génomes (extraits) de cette population seront croisés et mutés en vue de générer une nouvelle population qui sera ensuite classée. Ce processus est réitéré jusqu'à la non amélioration (stagnation) de la valeur d'adaptation (Goldberg, 1989).

3.1.2. Mise en oeuvre de l'algorithme génétique

Pour la mise en œuvre de l'algorithme génétique, nous avons codé les génomes qui représentent l'extrait moyennant le numéro de l'article et le numéro de la phrase dans cet article. L'ensemble suivant {(2,1), (1,3), (4,7), (6,6), (9,15), (0,0), (0,0)} illustre un génome formé par la première phrase du deuxième article, suivie de la troisième phrase du premier article, puis la septième phrase du quatrième article, etc. Le génome peut contenir des gènes non encore remplis par des phrases et qui sont présentés par le couple (0,0).

La population initiale des extraits, dont la taille est fixée à 10 est générée aléatoirement. Ainsi, il s'agit d'un choix arbitraire des gènes (phrases) qui vont constituer le génome (extrait). Les génomes ainsi constitués vont engendrer des descendants par des opérations de croisement et de mutation. Le croisement s'effectue de la façon suivante: le génome de chaque parent est coupé en une position tirée au hasard. Cette position est la même pour les deux génomes parents. La première séquence du premier génome parent va alors être collée à la seconde séquence du second génome parent et former ainsi un génome fils. Un second génome fils peut être formé de la même façon avec les deux parties des génomes restants. Le croisement est appliqué à chaque paire de génomes sélectionnée avec une probabilité fixée au préalable (la probabilité de croisement est égale à 0.7). Les génomes nouvellement engendrés vont être soumis à des mutations aléatoires qui vont modifier un ou plusieurs de leurs gènes (une ou plusieurs phrases). La mutation est appliquée avec une certaine probabilité (fixé à 0.1) aux génomes déjà croisés.

La population de génomes constituée (dont le nombre est égal à 200) est ensuite évaluée pour sélectionner les génomes « forts » qui vont former la population initiale de la prochaine génération (au nombre de 10). Pour quantifier la force d'un génome nous avons choisi d'évaluer ces génomes moyennant certains critères informationnels qui dépend de la teneur du génome en mots clés, sa longueur, et le poids de ses phrases (voir section suivante). Ce processus est réitéré jusqu'à ce qu'on détecte qu'il n'y a pas d'amélioration au fil d'un certain nombre de générations (30 générations).

Il est à noter que les paramètres de l'algorithme génétique, à savoir la taille de la population initiale et finale et les probabilités de croisement et de mutation, ont été déterminés suite à des expérimentations mettant en relief l'allure de convergence de l'algorithme avec les paramètres précités.

3.2. Critères d'évaluation et de classement des extraits

Pour évaluer l'extrait, nous avons retenu trois critères à savoir la longueur de l'extrait, sa couverture en mots clés et son poids (Jaoua et *al*., 2004). Dans un travail plus récent, Liu et *al*. ont introduit un quatrième critère d'anti-redondance dans le classement d'extraits pour l'élimination des phrases similaires (liu et *al*. 2006).

– Le critère de taille: ce critère permet de fixer la longueur de l'extrait en mots ou en phrases. Le coefficient d'importance ω_1 associé à ce critère est calculé comme suit:

Si $0{,}9 \times L_E \leq \sum_{i=1}^{m} L(S_i) \leq 1{,}1 \times L_E$ Alors ω_1=1

Si $\sum_{i=1}^{m} L(S_i) < 0{,}9 \times L_E$ Alors $\omega_1 = \frac{\sum_{i=1}^{m} L(S_i)}{L_E}$

Si $\sum_{i=1}^{m} L(S_i) > 1{,}1 \times L_E$ Alors $\omega_1 = 0$

$L(S_i)$: Longueur de la phrase i.

L_E : Longueur de l'extrait défini par l'utilisateur.

m : Nombre de phrases dans l'extrait.

– Le critère de couverture en mots clés : ce critère a pour objectif de garder les principaux mots ou concepts clés dans l'extrait final. En effet, un extrait est informatif s'il couvre tous les mots clés contenus dans les documents d'origine. Le coefficient d'importance ω_2 associé à ce critère est calculé comme suit :

$$\omega_2 = \frac{\sum M_{ext}}{\sum M_{doc}}$$

M_{ext} : Mot clé présent dans l'extrait.

M_{doc} **:** Mot clé présent dans le document.

Il est à noter qu'un mot est considéré comme étant un mot clé si sa fréquence dépasse trois fois le seuil moyen de fréquence des mots dans le groupe de documents.

– Le critère de pondération : ce critère préconise qu'un extrait doit contenir les phrases importantes des documents sources. Un résumé est important si le poids moyen des phrases qui le constituent est important. Le poids d'un terme est calculé en fonction de la valeur tf *idf . Le coefficient ω_3 asoocié à ce critère est calculé de la manière suivante.

$$\omega_3 = \frac{\sum tf(t) * idf(t)}{Nb}$$

tf : Fréquence du terme dans l'ensemble des textes sources.

idf : Nombre de documents du corpus/nombre de documents où apparaît le terme.

Nb : Nombre de termes dans la phrase.

Le classement des extraits dépend d'une fonction objectif qui permet d'agréger les critères précités. La fonction d'agrégation utilisée consiste à multiplier les valeurs associées à ces critères après être normalisés. Si l'extrait dépasse la taille fixée au préalable, sa valeur objectif est égale à zéro vu que le critère de longueur vaut zéro, ce qui permet d'éliminer les extraits dont la longueur excède celle désirée.

F(extrait) = $\tilde{\omega}_1 * \tilde{\omega}_2 * \tilde{\omega}_3$

$\tilde{\omega}_i$ Signifie que le coefficient associé au critère i est normalisé. Par exemple $\tilde{\omega}_3$, réfère au critère de pondération de l'extrait qui est normalisé en divisant ω_3 par le poids le plus élevé de l'extrait de la population.

4. Evaluation du système ExtraNews dans la compagne d'évaluation DUC

Plusieurs types d'évaluations ont été adoptés lors des conférences DUC pour quantifier les performances des systèmes de résumé automatique. Parmi ces évaluations nous citons l'évaluation ROUGE[2] : Recall-Oriented Understudy for Gistering Evaluation (Lin, 2004). Les mesures obtenues par le système Rouge sont générées d'une manière automatique et font intervenir la différence entre la distribution des mots (n_gram) d'un résumé candidat et celle d'un ensemble de résumés de référence. La formule de calcul des mesures Rouge est la suivante :

[2] Rouge : Recall-Oriented Understudy for Gistering Evaluation http://berouge.com

$$ROUGE_n = \frac{\sum_{C \in \{Référence\}} \sum_{n\_gram \in C} correspond(candidat, c)}{\sum_{C \in \{Référence\}} \sum_{n\_gram \in C}}$$

Où le terme *correspond (candidat,c)* représente le nombre maximum de N-grams communs entre le résumé système et le résumé de référence. Le dénominateur de l'équation est la somme du nombre de N-grams des résumés de références.

On note que $Rouge_n$ est la formule de base du score ROUGE et comme le montre l'équation précédente ce n'est que le paramètre du rappel. On peut donc obtenir des mesures de Rouge1 (1 gram), Rouge2 (2_gram), etc. Des études de corrélation ont montré que la mesure Rouge2 présente la meilleure corrélation avec les résumés humains (Lin, 2004). D'autres variantes ont été proposées et qui tiennent compte de l'ordre des mots ainsi que de la fonctionnalité grammaticale des n_gram (RougeBE).

Le système ExtraNews a participé dans les trois tâches qui s'intéressent aux résumés de documents multiples dans la session DUC'04 à savoir:

- La tâche 2 : consiste à résumer des documents multiples décrivant un évènement ;
- La tâche 4 : consiste à générer des résumés de documents traduits de la langue arabe ;
- La tâche 5 : consiste à produire des résumés de biographie humaine.

La même version du système ExtraNews a été expérimentée dans la DUC'07 et a participé à la tâche dédiée pour le résumé guidé par question utilisateur. Il est à noter que dans cette version, les mots de la question sont considérés comme des mots clés.

Nous nous limitons dans ce tableau à présenter les mesures Rouge2 obtenus par notre système dans les deux sessions d'évaluation.

	DUC'04 (identifiant : i=21, 23,24)			DUC'07 (id=28)
	Tâche 2	Tâche 4	Tâche 5	Principale
Rouge2	0.121	0.132	0.118	0.098
Rang /nombre des systèmes participants	4/14	1/11	3/14	16/32

Tableau 1. *Extrait des résultats de l'évaluation du système ExtraNews dans DUC'04 et DUC'07*

5. Améliorations apportées au système ExtraNews

Suite à l'examen de cette évaluation, nous avons constaté que la divergence des résultats obtenus lors des deux sessions DUC'04 et DUC'07 n'est pas due à la nature de la tâche évaluée, mais plutôt à la taille des documents en entrée. En effet, le nombre de documents par thème dans DUC'04 est égal à 10 avec une moyenne de 275 phrases par thème alors que dans DUC'07 il est de 25 documents regroupant en moyenne 720,1 phrases. Cette remarque corrèle avec l'expérience menée sur ce dernier corpus et qui a donné pour le système ExtraNews des résultats différents pour des exécutions distinctes sur un même thème. Ces résultats s'expliquent par la dérive génétique qui a caractérisé notre système lors de la conférence DUC'07 et qui est due au grand nombre de phrases traité en entrée. Cette dérive a été signalée par une altération de la fonction "objectif" utilisée dans le processus de classement des extraits des populations générées lors de l'application de l'algorithme génétique. Afin de corriger cette dérive, et d'améliorer la cohérence des extraits générés par le système ExtraNews, nous avons apporté trois types d'améliorations :

– La première amélioration consiste à ajouter un module de filtrage des phrases qui a pour objectif la minimisation du nombre de phrases en entrée. Ce module utilise la notion de dominance de phrases qui est une notion inspirée du domaine de l'ordonnancement multicritère. On dit qu'une phrase P domine la phrase Q si l'ensemble des mots clés de la phrase Q (ou de leurs synonymes[3]) est inclus dans l'ensemble des mots clés de la phrase P et si la longueur de la phrase P est inférieure ou égale à celle de la phrase Q. L'exemple suivant tiré du thème D0701A portant sur les grandes compensations des jugements des tribunaux américains illustre la dominance entre deux phrases.

Phrase P : ***Morris Dees**, co-founder of the **Southern Poverty** Law **Center** in **Montgomery**, Ala., represented the Keenans and has said he intends to take everything the **Aryan** Nations owns to **pay** the **judgment**, including the sect's **name**.*

Phrase Q: ***Morris Dees**, the co-founder of the **Southern Poverty** Law **Center** in **Montgomery**, Ala., and one of the attorneys for the plaintiffs, said he intended to enforce the **verdict**, taking everything the **Aryan** Nations owns, including its trademark **name**.*

L'ensemble des mots clés couvert par les deux phrases est composé des termes suivants : {*Dees, Southern, Morris, Poverty, Center, Montgomery, Aryan, Pay, judgment /verdict, name* }. La phrase P domine la phrase Q vu qu'elle couvre un mot clé de plus que la phrase Q (le mot *pay)* et que sa longueur est inférieure à celle de Q.

[3] Afin de chercher les synonymes des mots clés, nous avons utilisé le dictionnaire WordNet.

Le module de filtrage permet de déterminer les phrases dominées et de les éliminer permettant ainsi de réduire le pool des phrases initial utilisé pour générer des extraits. Il est à noter que la sélection des phrases dominantes permet, en outre, d'éliminer les phrases similaires ou synonymes. Il est à signaler que l'élimination des phrases synonymes dans l'extrait a été abordée comme critère de classification d'extraits dans les travaux de Liu (Liu et al, 2006).

– La deuxième amélioration consiste à apporter au processus d'exploration, utilisé par l'algorithme génétique, des mécanismes permettant d'éviter la dérive génétique. Il s'agit de définir une mémoire génétique permettant de sauvegarder au fil des générations les extraits potentiellement importants et de maintenir ainsi une diversité des solutions obtenues. On désigne par extraits potentiellement importants ceux qui ont éventuellement eu une fonction "objectif" meilleure que celles existantes dans la mémoire. Cette mémoire joue, donc, le rôle d'une mémoire qui regroupe les meilleurs extraits obtenus par l'algorithme génétique jusqu'à un moment donné. Il est à noter qu'en cas de dérive génétique c'est-à-dire de altération de la fonction "objectif", on retourne aux meilleurs extraits sauvegardés dans la mémoire pour former la population initiale et explorer ainsi, d'autres parties de l'espace de recherche.

– La troisième amélioration s'intéresse à la qualité de l'extrait produit dans la perspective de rétablir sa cohérence. Nous avons proposé, dans cette optique, l'intégration d'un module de révision dédié pour le contrôle et la résolution des problèmes d'incohérence rencontrés dans l'extrait final. Ces problèmes sont essentiellement manifestés par un désordre des phrases de cet extrait, par la présence d'expressions anaphoriques ou temporelles non résolues, etc. Jusqu'à présent, nous nous sommes focalisés sur la réorganisation des phrases du meilleur extrait sélectionné par notre système (Jaoua et *al*., 2008).

L'architecture de notre système, décrite dans la figure 1, distingue ainsi deux nouveaux modules à savoir le module de filtrage de phrases et celui de révision des extraits.

Après avoir intégré les deux modules précités, nous avons procédé à une évaluation de ExtraNews sur le même corpus utilisé lors de la session DUC'07. Les résultats obtenus montrent une baisse considérable du nombre des phrases en entrée de notre système (i.e. ce nombre a passé de 720 à 386 phrases en moyenne pour l'ensemble de documents formant un seul thème).

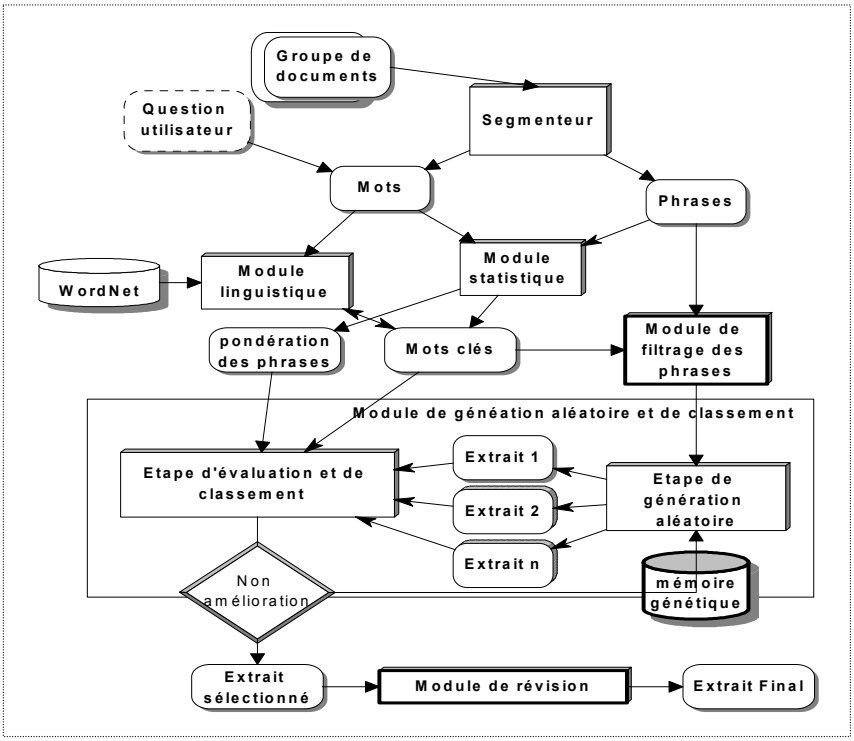

Figure 1. *Architecture du système ExtraNews*

Le tableau suivant rapportent les résultats obtenu après intégration de module de filtrage.

	DUC'04 (identifiant : i=21, 23,24)			DUC'07 (id=28)
	Tâche 2	Tâche 4	Tâche 5	Principale
Rouge2 initial	0.121	0.132	0.118	0.098
Rang /nb des systèmes	4/14	1/11	3/14	16/32
Rouge2 (avec filtrage)	0.122	0.132	0.120	0.118
Rang	3/14	1/11	3/14	4/32
Rouge2 (avec mémoire)	0.122	0.132	0.120	0.105
Rang	3/14	1/11	3/14	13/32
Rouge2 (avec mémoire et filtrage)	0.122	0.132	0.120	0.120
Rang	3/14	1/11	3/14	3/32

Tableau 2. *Evaluation des modules intégrés dans le système ExtraNews*

Les résultats de ce tableau montrent l'intérêt de l'intégration du module de filtrage, et de la mémoire génétique dans l'architecture de notre système. L'intégration du module de filtrage a pour but de faire face au nombre élevé de phrases des textes sources. Dans la même logique, l'intégration de la mémoire génétique a pour objectif de remédier à la dérive dans le cas où l'application de l'étape de filtrage ne permet pas une réduction importante du nombre de phrases initial.

6. Conclusion

Dans cet article, nous avons présenté les améliorations que nous avons apportées à notre système ExtraNews de résumé automatique de documents multiples. Les résultats obtenus lors de la session DUC'07 ont permis d'identifier l'importance du module de filtrage de phrases pour faire face à la dérive génétique qui a caractérisé l'algorithme génétique formant le noyau de notre système ExtraNews. La mise en œuvre de ce module s'est basée sur la notion de dominance des phrases en mots clés tout en tenant compte de leurs longueurs. L'intégration d'une mémoire génétique a permis aussi d'atténuer le problème de la dérive en récupérant les meilleures solutions en cas de convergence vers un maximum local.

L'examen des résultats obtenus montre aussi que l'on peut améliorer d'avantage les performances de notre système à travers l'application de nouveaux critères de sélection d'extraits. Dans ce cadre nous envisageons d'adopter d'autres critères de classements d'extraits après avoir étudié leur corrélation avec les indices d'évaluation utilisés par la compagne DUC. Cette évaluation doit, en outre, approfondir davantage les caractéristiques de la tâche de résumé à générer et donc de sélectionner les critères en fonction de la tâche proposée.

7. Bibliographie

Barzilay R., McKeown K., « Sentence Fusion for Multidocument News Summarization », *Computational linguistic,* V31, Mit Press, 2005, p. 297-327.

Brucker P, « On the complexity of clustering problems, in *Optimization and Operations Research », Lecture Notes in Economics and Mathematical Systems n° 157, M. Beckmann, H. Künzi (Eds.), Heidelberg, Springer-Verlag*, 1978, p. 45-54.

Carbonell J., Goldstein J., « The Use of MMR, Diversity-Based Reranking for Reordering Documents and Producing Summaries », *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Alistair Moffat and Justin Zobel, editors, 1998, Melbourne, Australia, p. 335-336.

Fellbaum C., *WordNet: An Electronic Lexical Database*, MIT Press, 1998.

Fuentes M., Massot M., Rodríguez H., Alonso L., « Headline extraction combining statistic and symbolic techniques », *Proceeding of DUC03*, Edmonton, Canada, Association for Computational Linguistics, May 31 - June 1, 2003.

Goldberg D.E., *Genetic algorithms in search, optimization, and machine learning*, Addison-Wesley, New York, 1989.

Jaoua K.F., Belguith H.L., Ben Hamadou A., «Révision des Extraits de Documents Multiples Basée sur la Réorganisation des Phrases », à paraître dans les actes du colloque IBIMA'08, 2008.

Jaoua K.F., Jaoua M., Belguith H.L., Ben Hamadou A., « Summarization at LARIS Laboratory », *Proceeding of the Document Understanding Workshop* Boston, USA May 6-7, 2004.

Jaoua M., Ben Hamadou A., «Automatic Text Summarization of Scientific Articles Based on Classification of Extract's Population», *Cicling'03*, 2003, p. 623-634.

Lin, C.Y. 2004, «Rouge: A Package for Automatic Evaluation of Summaries», *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*, 2004, p. 74-81.

Lin, C.Y., Hovy E., « Automated multi-document summarization in NeATS », *Proceedings of the DARPA Human Language Technology Conference,* 2002, p. 50–53.

Liu D., He Y., Ji D., Yang H., « Genetic Algorithm Based Multi-document Summarization », *PRICAI'06*, p.1140-1144.

Mani I., Bloedorn E., « Multi-Document Summarization by Graph Search and Matching », *Proceedings of the 14th National Conference on Artificial Intelligence*, 1997, Providence, Rhode Island, p. 622-628.

McKeown K., Hatzivassiloglou V., Klavans J.L., Holcombe M.L., Barzilay R., and Kan M.Y., « SIMFinder: A Flexible Clustering Tool for Summarization », *Proceedings of the Workshop on Automatic Summarization at the ACL,* 2001, p 41-49.

McKeown K. Klavans J., Hatzivassiloglou V., Barzilay R., and Eskin E., « Towards Multidocument Summarization by Reformulation: Progress and Prospects », *Proceedings of the 16th National Conference on Artificial Intelligence*, July 18-22, 1999, p. 453-460.

Minel J.L., *Filtrage sémantique,* Hermès , Paris, 2002.

Mori T., Nozawa M., and Asada Y., « Multi-Answer-Focused Multi-Document Summarization Using a Question-Answering Engine », *Proceedings of the 20th International Conference on Computational Linguistics (COLING 04)*, 2004, p. 439-445.

Over P., « An Introduction to DUC 2004 Intrinsic Evaluation of Generic New Text Summarization Systems », *Proceeding of the Document Understanding Workshop* Boston, USA May 6-7, 2004.

Radev D., McKeown K., « Generating natural language summaries from multiple on-line sources », *Proceeding of the Computational Linguistics*, V.24 (3), 1998, p. 469-500.

Saggion H., Gaizouskas R., « Multi-document summarization by cluster/profile relevance and redundancy removal », *Proceeding of the Document Understanding Workshop*, Boston, USA May 6-7, 2004.

White M., Korelsky T., Cardie C., Pierce D., Ng V., Wagstaff K., « Multidocument summarization via information extraction ». *Proceedings of the DARPA Human Language Technology Conference*, 2001, p. 143–146.