Évaluation de l'impact de l'intégration des étapes de filtrage et de compression dans le processus d'automatisation du résumé

Maher Jaoua[1], Fatma Kallel Jaoua[1], Lamia Hadrich Belguith[1] Abdelmajid Ben Hamadou[2]

1. Laboratoire MIRACL
Faculté des Sciences Economiques et de Gestion de Sfax
Route de l'aérodrome, BP 1088, 3018 Sfax, Tunisie
Maher.joaua@fsegs.rnu.tn, Fatma_fseg@yahoo.fr, l.belguith@fsegs.rnu.tn

2. Laboratoire MIRACL
Institut Supérieur d'Informatique et de Multimédia
Pôle technologique, Sakiet-Ezzit, BP 242, Sfax 3021, Tunisie
Abdelmajid.benhamadou@isimsf.rnu.tn

RÉSUMÉ. *Dans cet article, nous proposons une évaluation de l'impact de l'intégration des étapes de compression et de filtrage dans la chaîne de résumé automatique. Cette évaluation se base sur un certain nombre d'expériences que nous avons menées sur des sous-corpus disséminés lors la conférence DUC-TAC. Afin de mener ces expériences, nous avons adopté une méthode d'extraction qui considère le processus de résumé comme étant un problème d'optimisation où il s'agit d'en déterminer la meilleure partition qui répond à des critères prédéterminés. Les résultats obtenus montrent l'importance de l'intégration des étapes de filtrage et de compression.*

ABSTRACT. *We propose, in this article, the evaluation of the impact of the integration of compression and filtering steps in the automatic summarization process. This evaluation is based on a certain number of experiments that we have carried out on sub-corpora disseminated by the DUC-TAC conferences. In order to carry out these experiments, we have recommended a new extraction method which considers the summarization process as an optimization problem where it's a question to determine the most important partition that answers some predefined criteria. The obtained results show the importance of the integration of the filtering and compression steps.*

MOTS-CLÉS : *résumé automatique, filtrage de phrases, compression de phrases, évaluation des systèmes de résumés.*

KEYWORDS: *automatic summarization, sentences filtering, sentence compression, evaluation of summarization system.*

DOI:10.3166/DN.15.2.67-90 © 2012 Lavoisier

Document numérique – n° 2/2012, 67-90

1. Introduction

Plusieurs civilisations anciennes ont connu le résumé comme étant un texte plus réduit qui peut renseigner sur les idées importantes véhiculées par le document source. En effet, les Égyptiens à l'époque de la grande bibliothèque d'Alexandrie, ont utilisé des palettes d'argiles pour décrire le contenu des papyrus et ce dans l'objectif de les préserver. De même, la rédaction des résumés était une pratique courante et un style qui a caractérisé certains écrivains arabes lors de la traduction des œuvres grecs ; de tels résumés figurent, par exemple, dans le livre d'Averroès (1126-1198) *Talkhīs al-Āthār al-ʿulwiya* (Résumé des traces supérieures) qui résume certains écrits d'Aristote[1]. Les styles employés dans la rédaction de ces résumés étaient commentés par plusieurs philosophes arabes tels que al-Kalkashandi (1355-1418) qui a essayé, dans son ouvrage *Sobh al àcha fi sinàt al incha* (Le matin de l'hémérolope ou l'art de la rédaction), de décrire les étapes à entreprendre pour la traduction et la condensation des documents.

Dans la société contemporaine, le résumé est omniprésent dans la plupart des activités scientifiques et littéraires : il est couramment utilisé dans les articles scientifiques, les livres, les journaux, les médias, etc. De même, on retrouve l'épreuve de résumé dans les examens et les qualifications universitaires en raison de l'importance des compétences intellectuelles requises pour aborder les étapes de compréhension, de synthèse et de génération requises pour l'élaboration du résumé.

L'adoption par certains organismes d'une approche manuelle de résumé requiert des moyens de financement importants. Dans ce cadre, des services tels que le CSA « Chemical Service abstract » disposent de ressources humaines et matérielles pour résumer les articles dans le domaine de la chimie. Ces services transmettent ensuite les résumés générés manuellement aux chercheurs intéressés.

Compte tenu de la limite de l'approche manuelle, l'automatisation de l'activité résumante s'est imposée comme un défi auquel la communauté scientifique, actant dans le domaine de traitement automatique des langues naturelles, doit faire face. C'est pour cette raison que plusieurs paradigmes ont été entrepris pour aboutir à des méthodes automatiques visant à réduire le temps nécessaire pour générer le résumé par rapport à la tâche manuelle. L'ensemble de ces méthodes peut être répertorié selon le niveau d'analyse approprié. Ainsi, on peut distinguer les méthodes statistiques et les méthodes par analyse de surface. Ces méthodes optent pour la production d'extraits en se basant sur le repérage et l'extraction des phrases porteuses d'idées clés. En outre, les méthodes par analyse profonde consistent à produire un résumé en commençant par la construction d'une représentation interne de texte source et ce en se basant sur des formalismes sémantiques. Cette représentation sera ensuite réduite puis reformulée en résumé.

Dans cet article, nous mettons l'accent sur l'intérêt des étapes de compression et de filtrage dans le processus de résumé automatique. Nous présentons précisément un certain nombre d'expériences que nous avons menées sur un sous-ensemble des

1. Le mot *Talkhīs* est la traduction du mot résumé en arabe.

corpus disséminés lors de la conférence DUC[2]-TAC[3](DUC'04, TAC'08 et TAC'09). Le choix de ces corpus est motivé par la présence d'une évaluation manuelle qui permet de juger l'impact de l'intégration de ces étapes dans la chaîne de l'automatisation du résumé. Dans les deux sections suivantes, nous présentons les étapes du résumé ainsi que l'analogie que l'on peut dégager entre le processus automatique et manuel du résumé. Puis, nous effectuons un survol des principales méthodes développées dans le cadre de résumé de documents ainsi que des travaux réalisés dans le domaine de compression automatique. Ensuite, nous décrivons la méthode de base que nous avons adoptée pour le résumé automatique de texte. La méthode proposée constitue la plateforme permettant d'intégrer les étapes de compression et de filtrage et d'évaluer leur importance dans le processus du résumé. Nous enchaînons avec les expérimentations et les résultats obtenus.

2. Processus manuel du résumé

Les mécanismes cognitifs employés par l'être humain pour produire un résumé distinguent trois étapes. La première étape consiste à effectuer plusieurs lectures du texte source en vue de générer une représentation mentale permettant de détecter et de structurer les informations importantes du texte. Cette représentation peut être figurée par une carte cognitive qui retrace la hiérarchie des principaux thèmes abordés et les liens qui les relient. La deuxième étape consiste à réduire cette carte cognitive en se basant sur des mécanismes de suppression et de substitution, de sélection et d'intégration en vue de garder uniquement les thèmes intéressants. La troisième étape a pour objectif de réaliser le résumé. Elle s'appuie sur des mécanismes de génération qui dérivent des aptitudes linguistiques du résumé et qui ont pour rôle de reformuler la carte cognitive réduite en un texte bien formé. Il est à noter que les mécanismes employés dépendent particulièrement du type, de la taille et du style du résumé projeté.

Dans le cas où il s'agit de résumer une collection de documents, le processus humain de résumé est, en effet, plus complexe que celui de documents simples vu que les stratégies empruntées sont plus élaborées et nécessitent des recoupements inter-documents. Il s'agit, en l'occurrence, de comparer les informations propres à chaque document de la collection à résumer et d'en déduire les relations qu'ils entretiennent. Ainsi, la carte cognitive établie lors de la lecture du premier document sera augmentée de proche en proche par de nouveaux thèmes et de nouvelles associations au fur et à mesure qu'on consulte d'autres documents. Lors de la génération du résumé final, il est question de respecter la cohérence et la cohésion de l'ensemble tout en évitant les répétitions, les redondances et les discordances.

Le processus humain pour la production de résumé de documents simples ou multiples nécessite des compétences intellectuelles et linguistiques poussées. Cette remarque explique la faible qualité des résumés des débutants qui se contentent de

2. http://duc.nist.gov/
3. http://www.nist.gov/tac/

reproduire des phrases du ou des textes sources, et emploient généralement des mécanismes de paraphrasage pour rédiger leurs résumés.

3. Automatisation du résumé

L'adoption des outils de résumé automatique a pour but de pallier les insuffisances de l'approche manuelle très coûteuse en termes de temps, d'argent et d'effort. Bien que la qualité des résumés automatiques soit sujette à de nombreuses critiques en relation avec les discordances et les ruptures structurelles, le résumé automatique peut aider à juger l'importance du document ou de la collection de documents et si elle correspond aux attentes de l'utilisateur. La machine, grâce à sa puissance de calcul, peut traiter de grandes masses textuelles et faire les recoupements nécessaires en un temps réduit ; ce qui la place hors de concurrence humaine.

Le développement d'un système de résumé automatique doit cependant fixer auparavant, un certain nombre de facteurs qui dépendent des objectifs envisagés, du type, de la langue et du format de documents en entrée ainsi que des caractéristiques des résumés attendus. Jones Spärk (1999) détermine, dans ce contexte, trois types de facteurs qui régissent le développement des systèmes de résumé automatique :

– facteurs d'entrée : ces facteurs dépendent de la nature (texte, vidéo), de la langue (monolingue, multilingue), de la multiplicité (simple, multiple) et du genre (article de presse, article scientifique, monographie) des documents à résumer ;

– facteurs de but : ces facteurs décrivent le contexte d'énonciation du résumé tel que le type du résumé (extrait, abstrait), le niveau de spécialisation de l'audience, (spécialiste, novice) ainsi que l'objectif assigné par le résumé (assimilation, indexation, classification) ;

– facteurs de sortie : ces facteurs définissent la structure physique et logique du résumé à générer tels que la taille (résumé succinct, indicatif, informatif), la complétude (résumé général, ciblé) et la concision.

La prise en compte de ces facteurs dans la chaîne du résumé automatique peut être approchée de deux manières distinctes : la première adopte une analyse de surface alors que la seconde opère par analyse profonde. Le premier type d'analyses se base sur des techniques statistiques ou probabilistes. Les traitements, effectués par cette analyse, peuvent aussi être guidés par des indicateurs de surface. Ce type d'analyse est généralement emprunté par les méthodes de résumé automatique par extraction qui ont l'aptitude à traiter des documents relatifs à des domaines ouverts.

L'analyse est dite profonde si le processus de résumé automatique exploite des connaissances langagières dans l'objectif d'affiner les résultats du processus d'automatisation du résumé. La profondeur de l'analyse peut toucher plusieurs aspects de la langue. Par exemple, l'analyse de l'aspect lexical permet de détecter la correspondance entre les mots, alors que celle de l'aspect syntaxique détecte le rôle de chaque mot ou proposition dans la phrase. L'analyse sémantique permet, quant à elle, de construire une représentation interne de l'énoncé qui, après réduction, peut

être traduite en un abrégé. Il est à noter que ce type d'analyse ne peut être appliqué que pour des domaines restreints où la complexité de la modélisation des connaissances requises peut être maîtrisée.

Dans la section suivante, nous survolons les principaux travaux abordés dans le domaine de résumé automatique. Vu le nombre important de ces travaux, nous nous limitons à présenter ceux qui ont principalement marqué ce domaine.

4. Survol des travaux réalisés

Le nombre de travaux développés dans le domaine de résumé automatique est très important. Deux raisons interviennent dans cette synergie : la première résulte de la multiplicité des facteurs qui interviennent dans l'automatisation du résumé. La deuxième se manifeste par la création de plusieurs sessions d'évaluation telles que les conférences TIPSTER, DUC, MSE, TAC... Ces conférences d'évaluation à grande échelle offrent aux différents chercheurs la possibilité d'expérimenter leurs systèmes sur des grands corpus de documents. Il est à noter que ces conférences ont essayé de répondre aux différents facteurs d'entrée, d'objectif et de sortie. Parmi les facteurs d'entrée, nous distinguons la taille et le domaine des documents à condenser. En effet, nous rencontrons, dans la littérature, des méthodes de résumé automatique de documents simples, d'autres de documents multiples, celles appliquées à des domaines spécifiques, et d'autres plus générales. Un état plus complet figure dans les références suivantes : Minel, 2002 ; Spärk, 2007 ; Torres-Moreno 2011.

4.1. Survol des méthodes de résumé de documents simples

L'idée de produire des résumés automatiques n'est pas récente puisqu'elle date de la fin des années cinquante avec les travaux de Luhn (1958). Cependant, les techniques employées à cette époque ont été caractérisées par la simplicité de leurs traitements. Les méthodes actuelles peuvent être répertoriées selon trois axes : les méthodes statistiques, les méthodes par analyse de surfaces et celles par analyse profonde.

4.1.1. Méthodes statistiques

Ces méthodes sont plutôt orientées vers l'extraction des phrases clés des documents sources en se basant sur des critères et techniques d'ordre statistique. Les premiers travaux rencontrés, dans ce contexte, ont exploité des critères statistiques tels que la fréquence des mots dans le texte (Luhn, 1958), la position des phrases, la présence des mots dans la structure logique, ainsi que la présence d'expressions prototypiques (Edmundson, 1969). D'autres critères statistiques ont aussi été expérimentés pour déduire l'importance des phrases qui vont former l'extrait. Nous soulignons à titre d'exemple, la signature des mots qui a été proposée dans le cadre des travaux de Aone *et al.* (1999). D'autres techniques apparentées au domaine de l'extraction d'information se sont intéressées à la connectivité entre les phrases

(Mitra *et al.*, 1997), ainsi qu'à l'analyse sémantique latente pour déterminer l'importance des termes (Steinberger *et al.*, 2005).

Dans le cadre des méthodes statistiques, des techniques d'apprentissage ont aussi été expérimentées. Nous citons, à titre d'exemple, les travaux de Kupiec *et al.* (1995), qui s'appuient sur la classification bayésienne des phrases. Les paramètres de cette classification sont déterminés à partir d'un corpus annoté préalablement par des juges humains. D'autres techniques d'apprentissage ont également été expérimentées dans ce cadre telles que les modèles de Markov cachés (Conroy et O'leary, 2001), le support vecteur machine (Shen *et al.*, 2007) et les réseaux de neurones (Svore *et al.*, 2007).

4.1.2. Méthodes basées sur une analyse de surface

Ces méthodes exploitent le fait que certains indices linguistiques ou discursifs peuvent influencer l'importance des phrases. Dans ce cadre, les méthodes les plus représentatives sont celles fondées sur les « cue phrases » (Paice, 1990) ou celles qui exploitent certains indices pour attribuer des étiquettes sémantiques aux phrases (Berri, 1996 ; Minel et Desclés, 2000). De son côté, Lehmam (1995) repère des fragments textuels sur la base de scores calculés pour chaque phrase selon des termes préétablis. Certains travaux ont étudié la classification des phrases en fonction du rôle thématique qu'elles entretiennent dans le texte (Masson, 1998 ; Teufel et Moens, 1997 ; Ellouze, 2004). D'autres se sont basés sur les théories discursives telles que la RST pour en juger l'importance des propositions (Marcu, 1997) ou des phrases (Ono *et al.*, 1994 ; Maaloul et Keskes, 2010).

4.1.3. Méthodes basées sur une analyse profonde

Ces méthodes utilisent des modèles de connaissances en plus des modèles linguistiques afin de pouvoir déterminer et structurer les éléments essentiels qui figurent dans le texte source. Le premier type de méthodes se base sur la notion de templates (DeJong, 1982) qui reflètent des structures génériques qu'il s'agit de remplir par des informations extraites à partir du texte source. (Saggion et Lapalme, 2000). D'autres travaux utilisent les ontologies pour exploiter les connaissances dans un domaine particulier et par suite, sélectionner uniquement les éléments textuels qui véhiculent les concepts importants (Endres-Niggemeyer et Wansorra, 2004 ; Verma *et al.*, 2007).

4.1.4. Méthodes par compréhension

Les méthodes par compréhension sont aussi catégorisées parmi les méthodes par analyse profonde vu qu'elles opèrent par la génération d'une représentation interne du texte moyennant des analyses lexicales, syntaxiques et sémantiques. La génération du résumé passe par des étapes de réduction, de génération et de fusion (Mckeown *et al.*, 1995 ; Leskovec *et al.*, 2004).

4.2. Survol des méthodes de résumé de documents multiples

Les recherches sur le résumé automatique de documents multiples se distinguent par deux aspects : le premier réside dans le taux de réduction qui est assez élevé dans le cas des méthodes de résumé de documents multiples. Le second aspect est structurel vu que, dans le cas multiple, les systèmes de résumés doivent tenir compte des aspects de redondance, de cohérence, de cohésion et d'organisation des phrases du résumé ou de l'extrait produit. Pour la classification des méthodes actuelles, on peut adopter la même classification que celle du cas du résumé de documents simples.

4.2.1. Méthodes statistiques

La plupart des méthodes statistiques utilisent des métriques qui renseignent sur la similarité des unités textuelles des documents sources en vue de créer des groupes (clusters) ou d'éviter la redondance si le regroupement n'est pas effectué. Dans ce cadre, plusieurs méthodes ont expérimenté des métriques utilisées dans le cas du résumé de documents simples. En l'occurrence, nous citons les travaux de Daumé III *et al.* (2002), de Hardy *et al.* (2002) et de Boudin et Torres-Moreno (2007). D'autres chercheurs ont exploité des métriques spécifiques telles que la relevance marginale maximale (MMR) (Carbonell *et al.*, 1998 ; Goldstein *et al.*, 2000) ou le ratio de gain informationnel IGR (Mori *et al.*, 2004). D'autres travaux ont exploité la notion de centroïde introduite par Radev et qui désigne le vecteur des mots clés d'une collection de document (Radev *et al.*, 2001). La notion de centroïde a été reprise dans plusieurs travaux pour calculer les scores des phrases (Erkan et Radev, 2004) ou pour regrouper les phrases (Blair-Goldensohn *et al.*, 2006).

De nouvelles techniques sont aussi employées pour éviter la redondance. Nous citons, dans ce contexte, les travaux de Lacatusu *et al.* (2004) et de Barzilay et Lapata (2005) qui emploient des techniques d'alignement de séquences multiples rencontrées dans le domaine de la biologie. De son côté, Favre *et al.* (2006) utilisent des métriques de similarité d'Okapi, de Prosit et de SVM. Alors qu'Arora *et al.* utilisent la méthode d'allocation de Dirichlet latente (2008) qui se base sur la décomposition des documents en matrices.

Les techniques d'apprentissage ont aussi été utilisées pour le cas de résumé de documents multiples. Dans le contexte des travaux adoptant l'apprentissage supervisé, citons ceux de Conroy et Goldstein (2004) qui combinent la métrique de la relevance marginale maximale avec un modèle de Markov caché. Ce modèle utilise la signature de certains mots des documents en entrée, et ce en vue d'identifier les phrases importantes. Dans la même perspective, l'apprentissage a été utilisé pour inférer des mesures de similarité entre les phrases et les mots d'une requête utilisateur (Jagadeesh *et al.*, 2007). Alors que Amini et Usunier (2009) ont proposé une classification non supervisée pour décider de l'importance d'une classe de phrases par rapport aux mots de la question. De leur côté, Biadsy *et al.* (2008) ont exploité un classifieur pour décider de l'importance des phrases dans des textes de biographies.

4.2.2. Méthodes basées sur une analyse de surface

Les méthodes par analyse de surface exploitent des techniques linguistiques pour détecter les relations importantes qui peuvent s'établir entre les mots ou les phrases des documents sources. Dans ce cadre, Filatova *et al.* (2003) exploitent la relation entre les entités nommées pour déterminer l'importance de certains passages. Dans le même contexte, Nobata et Sekine (2004) se sont intéressés à classer les entités nommées en vue de déterminer l'importance des phrases. Zhou *et al.* (2004) ont aussi adopté une analyse de surface pour déterminer la catégorie des phrases issues d'articles décrivant la biographie d'une personnalité. Il est à noter que cette même stratégie a été adoptée par Ku *et al.* (2006) pour classer les polarités d'opinions dans les blogs.

L'analyse discursive, dans le cadre des résumés de documents multiples, permet d'explorer les relations rhétoriques qui existent entre les phrases et les documents sources. Ainsi, Radev *et al.*, (2002) ont proposé une extension de la théorie des structures rhétoriques appelée CST *(Cross Structure Theory)* pour modéliser les documents multiples. De son côté, Litowski (2003) a utilisé une représentation discursive décorée par les attributs syntaxiques et sémantiques des mots clés pour modéliser les phrases des documents sources.

4.2.3. Méthodes basées sur une analyse profonde

Parmi ces méthodes, nous distinguons celles qui proposent l'instanciation de templates. Citons, à titre d'exemple, les travaux de Radev *et al.*, (1998), White *et al.*, (2001) et Harabagiu *et al.* (2003). D'autres travaux ont exploité le paradigme de compréhension. Dans ce contexte, Mani et Bloedorn (1999) ont proposé une méthode de résumé de documents multiples basée sur la construction des graphes conceptuels de Sowa (1984). Dans un travail plus récent, Melli *et al.* (2006) ont proposé un système qui permet d'identifier les dépendances sémantiques entre les textes sources d'une part et la question de l'utilisateur d'autre part. D'autres travaux se sont intéressés à représenter la question afin de comparer cette représentation avec celle d'un texte source (Mohamed et Rajasekaran, 2006). Alors que Tennessy (2006) utilise une approche conceptuelle qui fait recours à une ontologie pour détecter les aspects temporels décrits dans les textes sources.

5. La compression des phrases

La compression de phrases s'est distinguée comme une étape importante permettant d'améliorer la qualité informationnelle du résumé ou de l'extrait. Elle constitue un maillon dans la chaîne de transition séparant le résumé de l'extrait. Le principe de la compression consiste à supprimer un ou plusieurs constituants d'une phrase dans le but de réduire le matériel textuel des phrases de la collection source. Les phrases compressées qui seront sélectionnées pour former le résumé auront, ainsi, une taille réduite ; ce qui va permettre d'ajouter de nouvelles phrases et donc d'améliorer la couverture globale de l'extrait pour une longueur fixée à priori. Notons que la compression peut considérer différents niveaux de granularité, c'est-

à-dire qu'il est possible d'éliminer un adjectif ou même une proposition entière (*e.g.* proposition relative, proposition subordonnée et proposition circonstancielle de lieu) à partir d'une phrase. Toutefois, deux éléments sont à prendre en compte lors de la compression de phrases : à savoir la grammaticalité et la concision. La grammaticalité a pour objectif de s'assurer que la phrase compressée respecte toujours les règles de la grammaire. Alors que la concision suppose qu'une phrase compressée doit retenir l'information essentielle de la phrase originale.

Les recherches élaborées dans le domaine de la compression automatique de phrases sont récentes et principalement développées en vue d'être intégrées dans les systèmes d'extraction et de résumé automatiques ou dans des systèmes de génération de titres. Selon qu'on utilise un modèle d'inférence ou non, nous pouvons classer les méthodes de compression suivant deux axes : les méthodes supervisées et les méthodes non supervisées.

Les méthodes supervisées préconisent l'utilisation de corpus parallèles qui permettent d'inférer des règles de compression. Citons, à titre d'exemple, les travaux de Knight *et al.* (2002) qui ont proposé une méthode de compression de phrases en langue anglaise qui se base sur la suppression, à partir des arbres syntaxiques construits, des éléments supposés non pertinents. La méthode utilise une technique probabiliste bayésienne qui est entraînée sur un corpus d'apprentissage. Il s'agit de déterminer la probabilité de suppression d'un syntagme sachant que la phrase résultat doit être bien formée. Dans le même contexte, citons aussi le travail développé par Nguyen *et al.* (2004) basé sur le modèle de Markov caché et qui a pour objectif de déduire la succession des règles de compression permettant de réduire la phrase. Dans le même cadre, Gagnon et Sylva (2006) ont proposé une méthode de compression de textes basée sur des filtres d'arbres syntaxiques en vue d'élaguer les parties non importantes de la phrase. Ces filtres sont déterminés à partir d'une étape d'apprentissage sur un corpus de test. Dans le même ordre d'idées, Clarke *et al.* (2010) ont exploité les arbres de décision en vue d'évaluer l'importance de chaque terme syntaxique et ce en utilisant des critères fonctionnels et discursifs. Parmi les méthodes de compression supervisées, citons aussi les travaux de Waszak *et al.* (2008) qui sont basés sur l'apprentissage d'un modèle formé de bigrammes et de trigrammes en utilisant un corpus aligné de phrases et de leurs formes compressées. Des critères entropiques sont utilisés pour le choix de la meilleure forme de compression possible. Cette même démarche a été empruntée par Fernández et Torres-Moreno (2009) qui utilisent un modèle issu des travaux sur la physique statistique pour représenter les mots de la phrase sous forme de spins. L'apprentissage s'effectue sur le couplage élaboré entre les différents spins pour déduire l'énergie textuelle des différentes formes de compression possibles.

Dans le cadre des méthodes non supervisées, Jing et McKeown (2000) ont proposé une méthode de compression qui se base sur la suppression des constituants jugés non pertinents par rapport au sujet principal du document, c'est-à-dire de ceux qui ne sont pas en rapport avec le thème du texte. Les auteurs utilisent une table statistique qui contient la probabilité de suppression de chaque constituant de la phrase. Dans la même classe de méthodes, nous pouvons intégrer les travaux de Knight et Marcu (2002) qui ont exploité un ensemble de règles permettant la

réécriture de l'arbre syntaxique de la phrase à compresser. Ces règles se basent sur des opérations élémentaires de transformation propres aux arbres syntaxiques. L'indisponibilité d'analyseurs syntaxiques dans certaines langues a poussé certains chercheurs à expérimenter d'autres niveaux d'analyse de la phrase. Dans ce contexte, McDonald (2006) a utilisé un parseur de dépendance pour repérer les relations entre les constituants élémentaires de la phrase tels que le sujet, verbe et complément. Ainsi, il a défini un ensemble de règles qui identifient, pour chaque couple de mots adjacents, ceux qu'on peut supprimer. D'un autre côté Cohn et Lapata (2009) ont définit la compression comme une tâche de réécriture utilisant des opérations de substitution, de réorganisation et d'insertion. Les mécanismes de réécriture opèrent sur des arbres et ont pour objectif de générer de nouveaux arbres plus compacts. Afin de limiter l'espace de réécriture, les auteurs utilisent une grammaire synchrone de substitution des arbres. Dans un travail récent, Yousfi Monod et Prince (2008) proposent la transformation des arbres syntaxiques en se basant sur des règles de réécriture déterminées manuellement. Le but de ces règles est de supprimer, à partir de la phrase, les éléments non gouverneurs dont la fonction syntaxique n'est pas importante.

6. Travaux élaborés dans le laboratoire MIRACL

Les travaux entrepris dans le laboratoire MIRACL[4] dans ce domaine de recherche se sont intéressés à différentes facettes du résumé. En effet, plusieurs améliorations ont été apportées à la méthode développée lors de la première participation du laboratoire à la conférence d'évaluation DUC en 2004 (Jaoua *et al.*, 2004). Ces participations ont été marquées par l'expérimentation de modules de compression (Jaoua, 2011), de filtrage et de réorganisation (Jaoua *et al.*, 2008). L'application pour la langue arabe a aussi été au centre des travaux du laboratoire dans ce domaine de recherche (Jaoua *et al.*, 2006 ; Maâloul *et al.*, 2010).

La synthèse de la majeure partie de ces travaux peut être élaborée en répondant à la question suivante : quel est l'impact de la compression et du filtrage dans le processus du résumé automatique ? Afin de répondre à cette question, nous commencçons par présenter le principe de base de la méthode adoptée pour l'automatisation du résumé ainsi que les extensions apportées. Nous décrivons le système ExtraNews bâti autour de cette méthode ainsi que les principaux modules mis en jeux, à savoir les modules de compression et de filtrage.

6.1. Principe de la méthode

La méthode que nous avons élaborée au sein du laboratoire MIRACL préconise que la production d'extrait doit se baser sur la comparaison d'un ensemble d'extraits pour sélectionner, ensuite, les meilleurs. Considérer l'extrait comme étant l'unité

4. MIRACL : Multimedia, InfoRmation systems and Advanced Computing Laboratory http://www.miracl.rnu.tn

minimale d'extraction suppose qu'on manipule plusieurs extraits à partir desquels il s'agit de sélectionner celui qui répond à certains critères. L'application de ces critères a pour objectif de maximiser la qualité de l'extrait tout en respectant d'autres paramètres tels que la taille de l'extrait. Ces propos peuvent être modélisés en tant que problème d'optimisation où il s'agit de choisir un sous-ensemble de phrases (une partition), parmi celles de la collection source, répondant à un certain nombre de critères. Ce problème peut être formalisé comme suit :

$$\begin{cases} \text{Choisir un sous-ensemble } \{xi\} \in \text{l'ensemble des phrases} \\ \text{des textes sources ;} \\ \text{Maximiser } F(xi) \end{cases} \quad (1)$$

avec i le numéro de la phrase et F(xi) une fonction qui dépend des critères définis sur l'extrait.

Soulignons que la détermination de l'ensemble des partitions d'un ensemble est un problème NP dur dont la complexité est un facteur exponentiel. La nature exponentielle de ce problème met en cause le choix d'une méthode exacte d'optimisation et nécessite, donc, l'utilisation de métaheuristiques. Ces dernières sont des méthodes d'optimisation approchées qui opèrent par l'évaluation de solutions intermédiaires dans l'objectif de converger vers des solutions meilleures. L'application de ces métaheuristiques pour le problème d'extraction suppose que toutes les solutions intermédiaires sont des extraits qui sont générés, en une première étape, puis évalués en fonction des critères utilisés. Notons que le processus de génération et d'évaluation nécessite plusieurs itérations pour améliorer les solutions sans toutefois garantir l'obtention de solutions optimales.

Parmi les méthodes d'optimisation approchées les plus utilisées, citons la méthode tabou, celle de recuit simulé et les algorithmes génétiques (AG). Nous avons opté pour l'utilisation des algorithmes génétiques dans l'objectif de générer des populations d'extraits intermédiaires à évaluer et classer. Ces algorithmes offrent l'avantage d'explorer une multitude de solutions en une seule itération, ce qui permet d'explorer un grand espace de recherche. Le deuxième argument en faveur de l'algorithme génétique réside dans sa simplicité et ses performances enregistrées dans de nombreux problèmes d'optimisation.

6.2. Présentation du système ExtraNews

Le schéma synoptique du système ExtraNews bâti autour de la méthode précitée est composé de six principaux modules dont l'interaction est illustrée dans la figure 1. Dans la suite, nous présentons un bref aperçu de l'ensemble des modules qui caractérisent l'architecture du système ExtraNews. Une description plus détaillée figure dans (Jaoua *et al.*, 2003 ; Jaoua, 2011).

– Le module de prétraitement : il permet d'extraire l'ensemble des phrases d'un document simple ou un ensemble de documents d'une collection source. Il permet

aussi de découper les phrases en mots en éliminant les balises et les DTD correspondants.

– Le module statistique : permet le calcul des fréquences des mots non outils ainsi que le tri de ces mots selon ces fréquences. Ce module tient compte des formes dérivées dans ce calcul. Les mots clés issus de ce module permettent de participer à certains critères d'évaluation des extraits tels que la couverture et la pertinence. Un mot est considéré comme mot clé si sa fréquence dans la collection s'écarte significativement de la normale.

– Le module linguistique : permet de corriger les fréquences des mots par la détection des mots synonymes et méronymes et ce en utilisant WordNet.

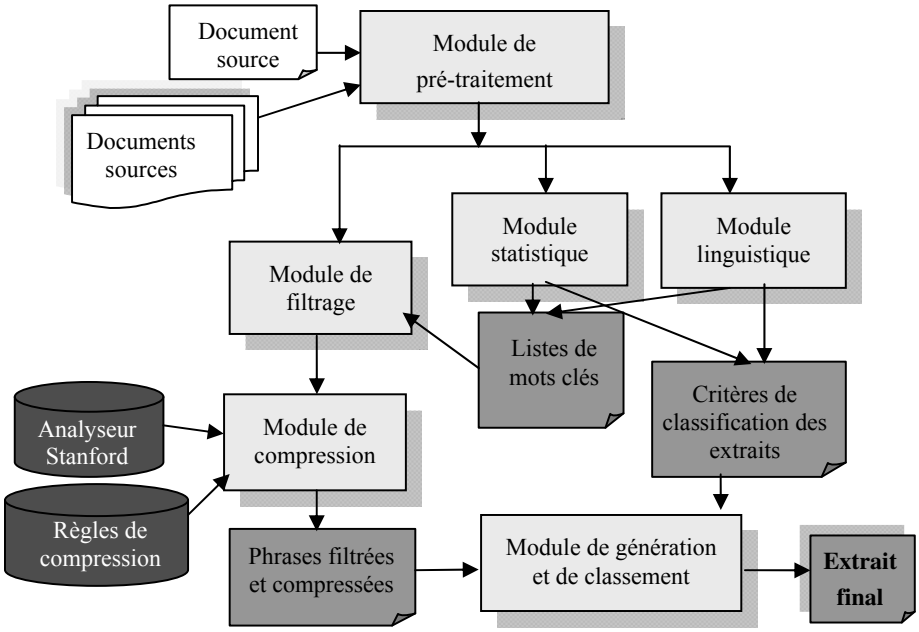

Figure 1. Schéma synoptique du système ExtraNews

– Le module de filtrage : permet de réduire les phrases en se basant sur la notion de dominance inter-phrastique. Ce module a pour tâche de réduire le nombre initial de phrases à manipuler en retenant uniquement celles non dominées.

– Le module de compression : permet de réduire le matériel textuel des phrases non dominées en éliminant certains constituants ne contenant pas des mots clés.

– Le module de génération et de classement : ce module s'intéresse à la génération et au classement des extraits moyennant une évaluation multicritère. Il utilise un algorithme génétique 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 qualité des extraits générés.

Dans ce travail, nous sommes particulièrement focalisés sur les modules de filtrage et de compression. C'est pour cette raison que nous détaillons ces deux modules, dans les deux sections suivantes.

6.3. Présentation du module de filtrage

L'étape de filtrage consiste à réduire le nombre de phrases de la collection des documents en entrée. Elle permet ainsi de réduire la complexité de l'étape de génération de partitions (d'extraits). Toutefois, puisque nous traitons des partitions, la suppression de phrases à partir du pool initial ne doit, en aucun cas, altérer le contenu informationnel des partitions qu'on peut générer.

Le filtrage des phrases peut se baser sur la notion de dominance que nous avons adoptée pour l'extrait et que nous adaptons pour la phrase comme suit.

Une phrase P domine une deuxième phrase Q si la taille de P est inférieure à celle de Q et l'ensemble des termes clés de la phrase Q est inclus dans la phrase P

Formellement, une phrase P domine une phrase Q et on note par Dom(P,Q) si :

$$L(P) < L(Q) \text{ et } Mots\_clés(Q) \subset Mots\_clés(P).$$

avec L(x) désigne la longueur de la phrases x ;

Mots_clés (x) désigne l'ensemble des mots clés de la phrase x.

Nous nous sommes intéressés particulièrement au critère de couverture en mots clés simples comme critère de filtrage des phrases. Ainsi, une phrase P domine une phrase Q si l'ensemble des mots clés de Q est inclus dans l'ensemble des mots clés de P et si la longueur de la phrase P est inférieure à celle de Q. L'étape de filtrage consiste ainsi à éliminer toutes les phrases dominées de la collection des documents sources et à ne garder que les phrases non dominées. L'exemple suivant, tiré de la collection D0701A de la conférence DUC'07, illustre le principe de 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 couverts 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 (le mot Pay) et que sa longueur est inférieure à celle de Q.

La notion de dominance peut être obtenue en s'intéressant à d'autres critères tels que la couverture en mots clés doubles ou l'ordre des mots clés. Nous avons intégré ces deux types de filtrages outre le filtrage précité que nous avons qualifié d'aléatoire vu qu'aucun ordre n'est exigé pour vérifier la dominance d'une phrase par apport à une autre phrase. Ainsi, nous avons introduit le filtrage ordonné qui exige la condition sur l'ordre d'apparition des mots pour juger la dominance. Alors que le filtrage double permet de vérifier la dominance des phrases en termes de mots clés doubles.

6.4. Description du module de compression

L'intégration du processus de compression dans notre contexte vise à augmenter les performances de la méthode proposée, notamment quand la taille de la collection est importante. Nous avons initialement appliqué ce processus sur le résumé pour le rendre succinct et aussi pour pouvoir ajouter à l'extrait de nouvelles phrases pour enrichir son contenu informationnel.

L'idée de base de la compression consiste à déterminer les constituants non essentiels des phrases et qui ne contiennent pas de mots clés. Notons que les mots clés peuvent regrouper les mots des titres, les mots clés issus de la question utilisateur et les mots clés simples dont la fréquence est importante (et qui ne sont pas des mots outils). En procédant de la sorte, nous minimisons les éventuelles pertes d'informations pertinentes engendrées par la compression. La nouvelle phrase compressée remplacera ainsi la phrase initiale.

Notons que la compression doit se baser sur une analyse syntaxique robuste permettant d'identifier et de distinguer les constituants essentiels du texte de ceux qui peuvent être supprimés. Dans ce contexte, nous avons utilisé l'analyseur de l'université de Stanford[5] en vue de déterminer les fonctions syntaxiques des différents constituants des phrases sources. Nous nous sommes aussi basés sur un ensemble de règles de compression que nous avons déterminées manuellement à partir d'un corpus de documents issu de la conférence DUC et qui ont été validées par des linguistes.

L'exemple ci-dessous détaille l'application de la compression pour la phrase suivante :

Phrase initiale: *The cardinals, who stopped granting on-the-record interviews more than a week ago, have offered few indications.*

5. http://nlp.stanford.edu/software/lex-parser.shtml

L'analyse syntaxique de cette phrase, en se basant sur l'analyseur de Stanford, génère l'arbre syntaxique suivant qui détaille pour chaque mot sa, fonction grammaticale selon le Penn Treebank :

« (ROOT (S (NP (NP (DT *The*) (NNS *cardinals*)) (, ,) (SBAR (WHNP (WP *who*)) (S (VP (VBD *stopped*) (S (VP (VBG *granting*) (NP (JJ *on-the-record*) (NNS *interviews*)) (NP (QP (JJR *more*) (IN *than*) (DT *a*)) (NN *week*) (NN *Ago*))))))) (, ,)) (VP (VBP *have*) (VP (VBN *offered*) (NP (JJ *few*) (NNS *indications*)))) (. .)))»

Étant donné la règle de compression suivante :

SPH1 +SBAR+SPH2 → SPH1+SPH2

L'application de cette règle sur la phrase précitée permet de supprimer la partie relative à SBAR (conjonction de subordination) et le résultat donne la phrase compressée suivante :

Forme compressée: *The cardinals have offered few indications*.

Notons de même que l'adjectif « few » peut aussi être éliminé en appliquant la règle suivante : VB +JJ +NN → VB+NN

Ainsi, la forme compressée après l'application de cette règle est la suivante :

The cardinals have offered indications.

Notons que les règles de compression ne s'appliquent que si la partie à supprimer ne contient pas de mots clés. Toutefois, une phrase peut se voir appliquer plusieurs règles de compression. Dans ce cas, nous continuons à trouver la forme compressée la plus réduite en termes de mots, tout en préservant la teneur en mots clés.

7. Expérimentations

Afin d'étudier l'impact du module de compression et de filtrage, nous avons modifié la fonction d'évaluation utilisée par le module de génération et de classement de notre système ExtraNews. Au lieu d'utiliser des critères maximisant la qualité informationnelle et structurelle de l'extrait, nous avons choisi comme fonction objectif une mesure intrinsèque qui évalue le degré de corrélation entre l'extrait et un ensemble de résumés de références. Parmi les mesures intrinsèques adoptées par les conférences d'évaluation, figure l'indice Rouge (Lin, 2004). L'indice $Rouge_n$ fait intervenir la correspondance entre la distribution des mots (n_grammes) d'un résumé candidat (PU) et celle d'un ensemble de résumés humains de référence (MU). La formule de calcul des mesures Rouge est la suivante :

$$Rouge_n = \frac{\sum_{C \in \{Référence\}} \sum_{n_{grammes} \in C} correspond\ (candidat, c)}{\sum_{C \in \{Référence\}} \sum_{n_{grammes} \in C}} \quad (2)$$

Où correspond (candidat, c) représente le nombre de n_grammes communs entre le résumé système et le résumé de référence. Le dénominateur de l'équation représente la somme du nombre de n_grammes des résumés de référence. Il est à noter que $Rouge_n$ est la formule générale du score Rouge. On peut donc obtenir des mesures de $Rouge_1$ (1_gramme), $Rouge_2$ (2_grammes), etc.

Des études récentes ont montré que la mesure $Rouge_2$ et $Rouge_{SU4}$ (bigrammes avec une fenêtre de 4 mots) présentent les meilleures corrélations avec les jugements humains (Liu et Liu, 2008). De ce fait, nous avons opté pour la mesure $Rouge_2$ pour quantifier l'importance d'un extrait. Ainsi, le système doit accepter non seulement le texte ou la collection à résumer, mais doit aussi accepter les résumés humains (appelés aussi résumés de références). La figure 2 retrace les modifications apportées au système ExtraNews lors de cette expérimentation.

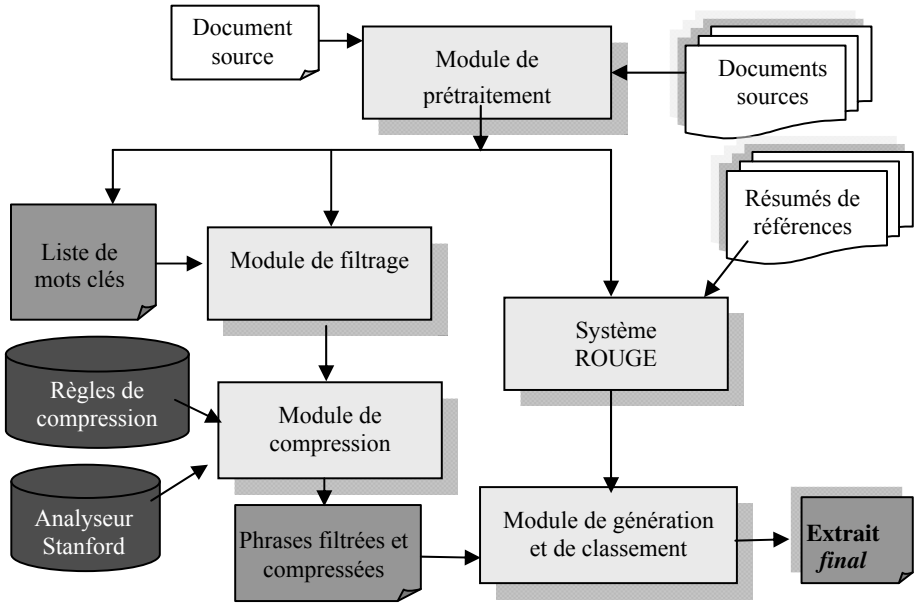

Figure 2. Modifications apportées au système ExtraNews pour l'expérimentation des modules de filtrage et de compression

Les corpus utilisés lors de ces expérimentations ont été extraits à partir de la conférence DUC'04 (Tâche 1) pour le cas de documents simples et des conférences DUC'04 (Tâche 5), DUC'08 (Tâche 1) et TAC'09 (Tâche 1) pour le cas de documents multiples. Notons que la tâche 1 de la conférence DUC'04 s'intéresse à générer des résumés succincts à partir de document simple. Alors que la tâche 5 de la même conférence s'intéresse à résumer une collection de documents correspondant à la biographie d'une personnalité. Pour la conférence TAC'08 et TAC'09 la tâche consiste à générer des résumés de documents multiples qui

présentent une évolution dans le temps. Le tableau 1 récapitule l'ensemble des corpus utilisés.

Tableau 1. Corpus utilisés

Corpus	Conférence	Nature	Nombre	Taille
C1	DUC'04 (Tâche 1)	Documents simples	50	75 caractères
C2	DUC'04 (Tâche 5)	Documents multiples (biographie)	30 (10 par collection)	665 caractères
C3	TAC'08 (Tâche 1)	Documents multiples	48 (10 par collection)	100 mots
C4	TAC'09 (Tâche 1)	Documents multiples	15 (10 par collection)	100 mots

Notre première expérience s'est intéressée à supprimer les modules de filtrage et de compression et à exécuter notre système sur les corpus précités. L'objectif étant de déterminer quelle est la valeur maximale (en $Rouge_2$) que peut atteindre un système d'extraction.

La deuxième expérience avait pour objectif d'évaluer le module de filtrage. Il s'agissait donc, pour le module d'évaluation, de traiter des phrases qui filtrées mais non compressées. Nous avons effectué cette expérience moyennant le filtrage aléatoire, ordonné et double. Dans la troisième expérience, nous avons utilisé les deux modules de filtrage avec ses différentes variantes conjointement avec le module de compression. Notons que l'ensemble des expériences ont été menées avec les mêmes conditions (mêmes paramètres et même nombre d'itérations pour l'algorithme de génération et de classement[6]).

8. Résultats obtenus

Les résultats de la première expérience, illustrée par le tableau 2, ont montré la présence d'une amélioration de l'indice Rouge qui a atteint, dans certains cas, la valeur de 39 % par rapport au meilleur système pour le cas de documents multiples (15 % pour le cas de document simple). Cette indication montre que les systèmes par extraction disposent encore d'une marge pour augmenter leurs résultats en améliorant le choix des phrases pertinentes (particulièrement pour le cas multiple).

6. Les paramètres de l'algorithme génétique utilisés lors de ces expériences sont :
Taille de la population initiale = 10 ; Taille de la population finale = 200 ; Taux de croisement = 0.65 ; Taux de mutation= 0.05 ; Fitness = $Rouge_2$; Critère d'arrêt : Stagnation.

Tableau 2. Résultats de la première expérience

Corpus	$Rouge_2$ (ExtraNews)	$Rouge_2$ (meilleur système)	$Rouge_2$ (ExtraNews) après expérience 1	Amélioration par rapport au meilleur système
C1	0.038	0.093	0.106	15%
C2	0.079	0.085	0.118	39%
C3	0.090	0.104	0.138	32%
C4	0.089	0.101	0.137	35%

Lors de la deuxième expérience, nous nous sommes intéressés à appliquer les trois types de filtrages à savoir : le filtrage aléatoire, le filtrage ordonné et le filtrage double. Les résultats obtenus, décrits dans le tableau 3, ont montré une légère amélioration des valeurs des indices $Rouge_2$ obtenues lorsqu'on applique le filtrage aléatoire.

Tableau 3. Impact de l'étape de filtrage

Corpus	$Rouge_2$ Expérience 1	Filtrage aléatoire	Filtrage ordonné	Filtrage double
C1	0.106	0.096	0.108	0.109
C2	0.118	0.123	0.112	0.114
C3	0.138	0.139	0.129	0.133
C4	0.137	0.137	0.130	0.131

Par contre, pour le filtrage ordonné ou double, nous avons remarqué une légère baisse des indices Rouge. Ces résultats peuvent être expliqués par le fait que le filtrage ordonné et double ne supprime pas un grand nombre de phrases, ce qui ne réduit pas l'espace de recherche du module de génération et de classement. À l'inverse et bien que le filtrage aléatoire puisse éventuellement supprimer certaines phrases ayant une corrélation importante avec les phrases des résumés de références, la réduction de l'espace de recherche permet de retrouver les résultats obtenus de la première expérience. Cette même remarque constitue la raison pour laquelle ce filtrage a obtenu l'indice Rouge le plus petit pour le cas de documents simples.

La troisième expérience, dont les résultats sont décrits dans le tableau 4, a montré une nette amélioration des indices Rouge par rapport à la première expérience. En effet, les phrases ont tendance à être courtes et le résumé dans ce cas peut englober d'autres phrases informatives qui peuvent corréler avec les résumés humains de référence. Le taux d'amélioration lors de cette expérience avoisine 20 % pour le cas de documents simples et de l'ordre de 10 % à 20 % pour le cas de documents multiples.

Tableau 4. Les améliorations moyennes issues de la combinaison des étapes de filtrage et de compression

Corpus	$Rouge_2$ expérience 1	Filtrage aléatoire + compression	Filtrage ordonné + compression	Filtrage double + compression	Amélioration
C1	0.106	0.121	0.129	0.129	21%
C2	0.118	0.141	0.133	0.136	19%
C3	0.138	0.152	0.139	0.141	10%
C4	0.137	0.148	0.141	0.139	08%

9. Conclusion

Les expériences que nous avons menées dans ce travail, ont mis l'accent sur l'intérêt que peut apporter l'intégration de chacune des étapes de filtrage et de compression dans le processus d'automatisation du résumé. Dans ce cadre, le filtrage aléatoire a enregistré les meilleurs résultats pour le résumé de documents multiples vu qu'il permet de réduire l'espace de recherche pour identifier les meilleurs extraits (en termes d'indice $Rouge_2$ après comparaison avec les résumés des références). La combinaison filtrage-compression a permis d'aboutir à des améliorations significatives (60 % pour le cas multiples et 38 % pour le cas simples). Ces valeurs permettent d'encourager les recherches dans ce contexte puisque les possibilités d'amélioration existent et elles sont prouvées par les expériences menées. Les résultats obtenus constituent en quelque sorte une limite expérimentale vu que les comparaisons obtenues ont été menées sur des résumés de références. Ce qui prouve que le champ d'amélioration pour les systèmes actuels est encore ouvert alors que la compression et le filtrage représentent l'une des dimensions de ce champ.

Comme perspective, nous pouvons réduire davantage les phrases sources et ce, en permettant au module de compression de générer plusieurs formes compressées pour la même phrase. Ainsi, nous gardons dans le même pool les phrases et les sous-phrases qui en découlent par compression. Dans ce cas particulier, nous devons envisager d'intégrer une autre étape de filtrage (post-filtrage), qui se place après le module de compression et qui a pour objectif de réduire le nombre de formes compressées en vue de diminuer l'espace de recherche. Les résultats escomptés seront probablement intéressants vu que les tailles des phrases seront plus réduites et que la probabilité d'intégrer d'autres phrases sera plus forte.

Une autre perspective à envisager consiste à évaluer les aspects linguistiques des résumés générés. En effet, le filtrage et la compression s'intéressent plutôt au contenu tout en négligeant les aspects linguistiques tels que la cohérence et la cohésion. En l'occurrence, la compression a tendance à dénuer la phrase de certains constituants essentiels qui peuvent nuire à la compréhension du résumé.

Bibliographie

Amini M.R., Usunier N. (2009). Incorporating prior knowledge into a transductive ranking algorithm for multi-document summarization. *Proceedings of the 32nd Annual International ACM Conference on Research and Development in Information Retrieval SIGIR'09*.

Aone C., Okurowski M.E., Gorlinsky J., Larsen B. (1999). A trainable summarizer with knowledge acquired from robust NLP techniques, *Advances in automatic text summarization*, MIT Press, p. 71-80.

Arora R., Ravindran R. (2008). Latent dirichlet allocation based multi-document summarization, *Proceedings of the Second Workshop on Analysis for Noisy Unstructured Data AND'08*.

Barzilay R., Lapata M. (2005). Collective content selection for concept-to-text generation, *Proceedings of Human Language Technology Conference HLT/EMNLP'05*.

Berri A. (1996). *Contribution à la méthode d'exploration contextuelle; applications au résumé automatique et aux représentations temporelles; réalisation informatique du système SERAPHIN*, Thèse de doctorat, Université de Paris-Sorbonne I.S.H.A.

Biadsy F., Hirschberg J., Filatova E. (2008). An unsupervised approach to biography production using Wikipedia. *Proceedings of the Association meeting of Computational Linguistics ACL'08*, p. 807-815.

Blair-Goldensohn S., McKeown K. (2006). Integrating rhetorical-semantic relation models for query-focused summarization. *Proceedings of the Document Understanding Conference DUC'06*.

Boudin F., Torres-Moreno J.M. (2007). NEO-CORTEX: a performant user-oriented multi-document summarization system. *Proceedings of Computational Linguistics and Intelligent Text Processing CICLing'07*, p. 551-562.

Carbonell J., Goldstein J. (1998). 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*, p. 335-336.

Clarke J., Lapata M. (2010). Discourse constraints for document compression. *Journal of Computational Linguistics*, vol. 36, n°3, p. 411-441.

Cohn T., Lapata M. (2009). Sentence compression as tree transduction. *Journal of Artificial Intelligence Research JAIR*, vol. 34, n°1, p. 637-674.

Conroy J.M., Goldstein J., Schlesinger J.D., O'leary D.P. (2004). Left-brain/right-brain multi-document summarization, *Proceedings of Document Understanding Conference DUC'04*.

Conroy J.M., O'leary D.P. (2001). Text summarization via hidden Markov models, *Proceedigs of the 24th annual international ACM SIGIR conference on research and development in Information Retrieval*, p. 406-407.

Daumé III H., Echihabi A., Marcu D., Munteanu D.S., Soricu R. (2002). GLEANS: A Generator of Logical Extracts and Abstracts for Nice Summaries, *Proceedings of the Document Understanding Conference DUC'02*, Philadelphia, PA.

DeJong G. (1982). An overview of the FRUMP system, *Strategies for natural language processing*, W.G. Lehnert and M.H. Ringle (ed.), Hillsdale, Erlbaum, p.149-176.

Edmundson H.P. (1969). New methods in automatic abstracting, *Journal of the ACM*, vol. 16, n°2, p. 264-285.

Ellouze M. (2004). *Utilisation des schémas rhétoriques pour le contrôle de la cohérence et génération de résumes automatiques d'articles scientifiques*, Thèse de doctorat, École Nationale des Sciences de l'Informatique, Université de Manouba.

Endres-Niggemeyer B., Wansorra E. (2004). Making cognitive summarization agents work in a real-world domain, *Proceedings of NLUCS Workshop*, p. 86-96.

Erkan G., Radev D. (2004). LexPageRank: prestige in multidocument text summarization. *Proceedings of Empirical Methods in Natural Language Processing EMNLP'04*, p. 465-471.

Favre B., Béchet F., Bellot P., Boudin F., El-Bèze M., Gillard L., Lapalme G., Torres-Moreno J.M. (2006). The LIA-Thales summarization system at DUC-2006, *Proceedings of Document Understanding Conference DUC'06*, New York, USA.

Fernández S., Torres-Moreno J.M. (2009). Une approche exploratoire de compression automatique de phrases basée sur des critères thermodynamiques. *Actes de la conférence Traitement Automatique des Langues Naturelles : TALN'2009*, Senlis, France.

Filatova E., Hatzivassiloglou V. (2003). Domain-independent detection, extraction, and labelling of atomic events. *Proceedings of the Recent Advances in Natural Language Processing RANLP'03*.

Gagnon M., Sylva L.D. (2006). Text compression by syntactic pruning, *Proceedings of the 19th Conference of the Canadian Society for Computational Studies of Intelligence AI'06*, p. 312-323.

Goldstein J., Mittal V., Carbonell, J., Callan J. (2000). Creating and evaluating multidocument sentence extract summaries, *Proceedings of the 9th International Conference on Informaiton and Knowledge Management*, p. 165-172.

Harabagiu S.M., Lacatusu F.V., Maiorano S.J. (2003). Multi-document summaries based on semantic redundancy. *Proceedings of the 14th Florida Artificial Intelligence Conference FLAIRS'03*, p. 387-391.

Hardy H., Shimizu N., Strzalkowski T., Ting L., Zhang X., Wise G.B. (2002). Cross-document summarization by concept classification, *Proceedings of the 25th Annual International Conference on Research and Development in Information Retrieval SIGIR'02*, p. 121-128.

Jagadeesh J., Prasad P., Vasudeva V. (2007). Capturing sentence prior for query-based multi-document summarization, *Proceedings of the conference: Recherche d'Information Assisté par Ordinateur RIAO'07*.

Jaoua K.F. (2011). *Une méthode flexible de résumé automatiques de documents multiples*. Thèse de doctorat, Faculté des sciences économiques et de gestion de Sfax, Université de Sfax, Tunisie.

Jaoua K.F., Belguith H.L., Jaoua M., Ben Hamadou A. (2008). Intégration d'une étape de pré-filtrage et d'une fonction multi-objectif en vue d'améliorer le système ExtraNews de résumé de documents multiples. *Actes de la conférence Traitement Automatique des Langues Naturelles TALN'08*.

Jaoua K.F., Jaoua M., Ellouze M., Maaloul M.H., Belguith H.L., Ben Hamadou A. (2006). Vers le résumé automatique de documents arabes. *Actes de l'atelier SIA : Systèmes d'Information Arabisés, en marge du congrès INFORSID'06*.

Jaoua K.F., Jaoua M., Ben Hamadou A. (2004). Summarization at LARIS laboratory, *Proceedings of Document Understanding Conference DUC'04*.

Jaoua M., Ben Hamadou A. (2003). Automatic text summarization of scientific articles based on classification of extract's population. *Proceedings of the Conference on Intelligent Text Processing and Computational Linguistics CICLing'03*.

Jing H., McKeown K. (2000). Cut and paste based text summarization. *Proceedings of the 1st Meeting of the North American Chapter of the Association for Computational Linguistics NAACL'00*, p. 178-185.

Knight K., Marcu D. (2002). Summarization beyond sentence extraction: a probabilistic approach to sentence compression. *Journal of Artificial Intelligence*, vol. 139, n° 1, p. 91-107, 2002.

Ku L.W., Liang Y.T., Chen H.H. (2006). Opinion extraction, summarization and tracking in news and blog corpora. *Proceedings of AAAI Symposium on Computational Approaches to Analysing Weblog*.

Kupiec J., Pedersen J., Chen F. (1995). A trainable document summarizer, *Proceedings of the 18th Annual International ACM Conference on Research and Development in Information Retrieval SIGIR'95*, p. 68-73.

Lacatusu F., Maiorano S., Harabagiu S. (2004). Multi-Document summarization using multiple-sequence alignment. *Proceedings of the International Conference on Language Resources and Evaluation LREC'04*.

Lehmam A. (1995). *Le résumé de textes techniques et scientifiques, aspects linguistiques et computationnels.* Thèse de doctorat, Université de Nancy 2.

Leskovec J., Grobelnik M., Milic-Frayling N. (2004). Learning sub-structures of document semantic graphs for document summarization. *Proceedings of KDD* Workshop on *Link Analysis and Group Detection LinkKDD*.

Lin C.Y. (2004). ROUGE: A package for automatic evaluation of summaries. *Proceedings of the Workshop on Text Summarization Branches Out: WAS'04*.

Litowski, K.C. (2003). Text summarization using xml-tagged documents, *Proceedings of the Document Understanding Conference DUC'03*, Edmonton, Canada.

Liu F., Liu Y. (2008). Correlation between ROUGE and human evaluation of extractive meeting summaries. *Proceedings of ACL/HLT'08*, p. 201-204.

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

Maâloul M.H., Keskes I. (2010). Résumé automatique de documents arabes basé sur la technique RST, *Actes de la Rencontre des Etudiant Chercheurs en Informatique pour le Traitement Automatique des Langue RECITAL*, Montréal, Canada.

Mani I., Bloedorn E. (1999). Summarizing similarities and differences among related documents, *Information Retrieval*, p. 1-23.

Marcu D. (1997). From discourse structures to text summaries. *Proceedings of ACL Workshop on Intelligent Scalable Text Summarization*, p.82-88.

Masson N. (1998). *Méthodes pour une génération variable de résumé automatique : Vers un système de réduction de textes*, Thèse de doctorat, Université Paris-11.

McDonald R. (2006). Discriminative sentence compression with soft syntactic evidence. *Proceedings of the 11th meeting of EACL'06*, p. 297-304.

McKeown K., Robin J., Kukich K. (1995). Designing and evaluating a new revision-based model for summary generation, *Information Processing and Management*, vol. 3, n° 5.

Melli G., Shi Z., Wang Y., Liu Y., Sarkar A., Popowich F. (2006). Description of SQUASH, the SFU question answering summary handler for the DUC-2006 summarization task, *Proceedings of the Document Understanding Conference DUC'06 at HLT/NAACL'06*, New York.

Minel J.L. (2002). *Filtrage Sémantique : Du Résumé automatique à la fouille de textes,* Hermès.

Minel J.L., Desclés J.P. (2000). Résumé automatique et filtrage des textes, *Ingénierie des langues*, J-M. Pierrel (éditeur), Paris, Hermès, p. 253-270.

Mitra M., Singhal A., Buckley C. (1997). Automatic text summarization by paragraph extraction, *Proceedings of the ACL'97/EACL'97 Workshop on Intelligent Scalable Text Summarization*, p. 31-36.

Mohamed A., Rajasekaran S. (2006). Improving query-based summarization using document graphs. *Proceedings of the IEEE International Symposium on Signal Processing and Information Technology*, p. 408-410.

Mori T., Nozawa M., Asada Y. (2004). Multi-answer-focused multi-document summarization using a question-answering engine. *Proceedings of the 20th International Conference on Computational Linguistics COLING'04*, p. 439-445.

Nguyen M.L., Shimazu A., Horiguchi S., Ho T.B., Fukushi M. (2004). Probabilistic sentence reduction using support vector machines, *Proceedings of the 20th international conference on Computational Linguistics COLING'04*, p. 743-749, Geneva, Switzerland.

Nobata C., Sekine S. (2004). CRL/NYU summarization system at DUC 2004, *Proceedings of Document Understanding Conference DUC'04*.

Ono K., Sumita, K., Miike S. (1994). Abstract generation based on rhetorical structure extraction, *Proceedings of COLING'94*, p. 344-348.

Paice C.D. (1990). Constructing literature abstracts by computer: techniques and prospects, *Information processing management*, vol. 26, n°1, p. 171-186.

Radev D., Zhang Z., Blair-Goldensohn S. (2002). Towards CST-enhanced summarization, *Proceedings of AAAI/IAAI'02*, p. 439-446.

Radev D., Blair-Goldensohn S., Zhang Z. (2001). Experiments in single and multidocument summarization using MEAD. *Proceedings of the Document Understanding Conference DUC'01*.

Radev D., Mckeown K. (1998). Generating natural language summaries from multiple on-line sources. *Journal of Computational Linguistics,* vol. 24, n° 3, p. 469-500.

Saggion H., Lapalme G. (2000). Selective analysis for the automatic abstracting: evaluating indicativeness and acceptability, *Proceedings of Content-Based Multimedia Information Access, RIAO'00*, p. 747-764, France.

Shen D., Sun J-T., Li H., Yang O., Chen Z (2007). Document summarization using conditional random fields, *Proceedings of IJCAI'07*, p. 2862-2867.

Sowa J.F. (1984). *Conceptual Structures: Information Processing in Mind and Machines*. Addison-Wesley, Reading (MA), USA.

Spärck Jones K. (2007). Automatic summarising: The state of the art, *Information Processing and Management,* vol. 43, n° 6, p. 1449-1481.

Spärck Jones K. (1999). Automatic summarising: Factors and directions. *Advances in Automatic Text Summarizatio*n, MIT Press, p. 1-12.

Steinberger J., Ježek K. (2005). Text summarization and singular value decomposition, *Proceedings of Advances in Information Systems,* p. 245-254.

Svore K.M., Vanderwende L., Burges C.J.C. (2007). Enhancing single-document summarization by combining ranknet and third-party sources, *Proceedings Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, p. 448-457.

Tennessy B. (2006). *An Epistemological approach to domain-specific multiple biographical document summarization*. PhD Thesis, University of British Columbia, USA.

Teufel S., Moens M. (1997). Sentence extraction as a classification task, *Proceedings of the Workshop on Intelligent Scalable Text Summarization at the ACL/EACL conference*, Madrid, Spain.

Torres-Moreno J.M. (2011). *Résumé automatique de documents - une approche statistique*, Editions Hermès-Lavoisier.

Verma R., Chen P., Lu W. (2007). A semantic free-text summarization system using ontology knowledge, *Proceedings of the Document Understanding Conference DUC'07*.

Waszak T., Torres-Moreno J.M. (2008). Compression entropique de phrases contrôlée par un perceptron. *Actes des Journées internationales d'Analyse statistique des Données Textuelles JADT'08*, vol. 2, p. 1163-1173.

White M., Korelsky T., Cardie C., Ng V., Pierce D., Wagstaff K. (2001). Multidocument summarization via information extraction, *Proceedings of the Document Understanding Conference DUC'01*, p. 143-146.

Yousfi Monod M., Prince V. (2008). Sentence compression as a step in summarization or an alternative path in text shortening. *Proceedings of the 24th International Conference on Computational Linguistics COLING'08*, p. 139-142.

Zhou L., Ticrea M., Hovy H. (2004). Multi-document biography summarization, *Proceedings of the conference on Empirical Methods in Natural Language Processing, EMNLP'04*, p. 434-441, Barcelona, Spain.