Filtrage pour la construction de résumés multi-documents guidée par un profil

Olivier Ferret, Sana Leila Châar, Christian Fluhr

Filtrage pour la construction de résumés multi-documents guidée par un profil

Sana Leila Châar, Olivier Ferret et Christian Fluhr

CEA – LIST/LIC2M (Laboratoire d'Ingénierie de la Connaissance Multimédia Multilingue)
18, route du Panorama
BP 6
92265 Fontenay-aux-Roses Cedex
[chaars, ferreto, fluhrc]@zoe.cea.fr

RÉSUMÉ. *Dans cet article, nous présentons une méthode de filtrage permettant de sélectionner à partir d'un ensemble de documents les extraits de textes les plus significatifs relativement à un profil défini par un utilisateur. Pour ce faire, nous mettons l'accent sur l'utilisation conjointe de profils structurés et d'une analyse thématique des documents. Cette analyse permet également d'étendre le vocabulaire définissant un profil en fonction du document traité en sélectionnant les termes de ce dernier les plus étroitement liés aux termes du profil. Tous ces aspects assurent une plus grande finesse du filtrage tout en permettant la sélection d'extraits de documents ayant un lien plus ténu avec les profils mais davantage susceptibles d'apporter des informations nouvelles et donc intéressantes. L'intérêt de l'approche présentée a été illustré au travers du système REDUIT qui a fait l'objet d'une évaluation concernant à la fois le filtrage de documents et l'extraction de passages.*

ABSTRACT. *In this article, we present an information filtering method that selects from a set of documents their most significant excerpts in relation to an user profile. This method relies on both structured profiles and a topical analysis of documents. The topical analysis is also used for expanding a profile in relation to a particular document by selecting the terms of the document that are closely linked to those of the profile. This expansion is a way for selecting in a more reliable way excerpts that are linked to profiles but also for selecting excerpts that may bring new and interesting information about their topics. This method was implemented by the REDUIT system, which was successfully evaluated for document filtering and passage extraction.*

MOTS-CLÉS : *Filtrage d'information, profil utilisateur, résumé multi-document.*

KEYWORDS: *Information filtering, user profile, multi-document summarization.*

1. Introduction

Dans un souci de se rapprocher des besoins des utilisateurs confrontés à une masse toujours croissante d'informations sous forme électronique, le domaine de la recherche d'information s'est élargi ces dernières années en direction d'applications visant non plus seulement à retrouver des documents à partir d'un ensemble de mots-clés mais à rechercher l'information répondant à la demande d'un utilisateur, demande pouvant s'exprimer sous des formes diverses. Le succès actuel de la problématique des systèmes de question/réponse, incarnée par l'évaluation Question/Answering de TREC (Voorhees, 2000), en est une illustration. Lorsque la question posée est de nature factuelle, il est possible de fournir un court extrait de document contenant la réponse attendue. En revanche, lorsque la question est plutôt de nature thématique, une réponse ne peut être obtenue qu'en rassemblant, en recoupant et en compilant un ensemble d'informations issues de différents documents. Le travail que nous présentons dans cet article s'inscrit dans cette seconde perspective. Sa problématique rejoint donc celle du résumé multi-document, faisant actuellement l'objet de nombreux travaux, en particulier dans le cadre de l'évaluation DUC (Document Understanding Conference) (Over, 2001).

La plupart des systèmes de résumé, qu'ils soient mono ou multi-document, fonctionnent actuellement par extraction de passages ou de phrases. Différentes approches ont été développées au sein de ce paradigme. L'une des plus répandues consiste à exploiter des critères essentiellement statistiques ou probabilistes. C'est le cas de la méthode MMR-MD (Goldstein *et al.*, 2000), qui définit une métrique d'intérêt des passages de texte, du système développé par TNO-TPD (Kraaij *et al.*, 2001), qui attribue un score à des segments textuels en combinant un modèle de langage unigramme et un modèle bayésien, ou encore des systèmes décrits dans (Boros *et al.*, 2001) et (Stein *et al.*, 2001), qui font appel à des méthodes de classification.

L'exploitation de critères plus linguistiques est une deuxième grande approche pour le résumé automatique. Elle est particulièrement représentée pour les systèmes de résumé mono-document par le système de Lehmam (Lehmam, 1995), les travaux de Minel (Minel, 2000) s'appuyant sur la plate-forme ContextO et le système SumUm de Saggion (Saggion *et al.*, 2002). Le premier s'appuie sur le repérage de marqueurs de surface dans les phrases d'un document pour déterminer si une phrase est à retenir pour la construction du résumé. L'originalité du système ContextO est quant à lui de donner les moyens d'accéder au contenu sémantique du document par une analyse linguistique des documents en utilisant la méthode d'exploration contextuelle (Desclés, 1987) pour ensuite en extraire certaines séquences particulièrement pertinentes. Enfin, le système SumUM cherche lui aussi à accéder au contenu conceptuel des documents par le biais de plusieurs analyses linguistiques (extraction de termes, patrons de sélection des contenus informatifs) mais dans une perspective plus proche de l'extraction d'information.

Dans le domaine des systèmes de résumé multi-document, l'approche « linguistique » est typiquement représentée par le travail de Radev *et al.* (1998). Les systèmes s'inscrivant dans cette perspective font appel à des traitements linguistiques plus ou moins élaborés (extraction de termes, reconnaissance d'entités nommées, analyse syntaxique, …) pouvant aller dans certains cas jusqu'à la reformulation des phrases extraites grâce à un module de génération, comme par exemple dans les systèmes développés par Barzilay (Barzilay *et al.*, 1999) ou McKeown (McKeown *et al.*, 2001). Dans cette même perspective, des travaux tels que (Mani *et al.*, 1999) ou (White *et al.*, 2001) utilisent des méthodes d'extraction d'information et de traitement linguistique pour produire des résumés multi-documents en fonction d'une requête ou d'un profil utilisateur.

Le travail que nous présentons dans cet article vise pour sa part à donner à un utilisateur la possibilité de parcourir rapidement un ensemble de documents selon un point de vue particulier, par exemple à la suite d'une requête réalisée auprès d'un moteur de recherche. Ce point de vue est représenté par le biais d'un profil. Le parcours se fonde quant à lui sur l'extraction des passages les plus étroitement en relation avec ce profil en s'appuyant sur un traitement linguistique et une analyse thématique des documents.

Nous commencerons dans la section 2 par une description plus ample des différentes approches que nous venons de citer puis nous présenterons dans la section 3 les choix que nous avons adoptés ainsi que l'architecture du système REDUIT[1] qui les implémente et ses différents composants. La section 4 sera consacrée à la définition de la notion de profil utilisateur ainsi qu'à la description d'une méthode pour les structurer automatiquement. Puis nous exposerons la technique d'analyse des documents que nous utilisons au cours de la section 5. L'exposé de notre méthode de filtrage sera l'objet de la section 6, nous présenterons dans la section 7 les règles utilisées pour fusionner l'information retenue pour la construction de résumé et dans la section 8, les règles prises en compte pour construire le résumé final. La section 9 sera consacrée à la méthode d'évaluation que nous avons retenue pour ce travail ainsi qu'aux résultats obtenus par REDUIT dans ce cadre. Enfin pour clore cet article, nous évoquerons quelques pistes possibles pour la continuation de ce travail.

2. Analyse des approches qui ont influencé notre démarche

2.1. *Les approches fondées sur un traitement purement numérique*

Les approches dites numériques se fondent sur la mise en place de métriques permettant de délimiter dans les documents à résumer les extraits à sélectionner pour la construction du résumé. Les travaux de Goldstein sur la méthode MMR-MD (Marginal Relevance Multi-Document) illustrent parfaitement ce type d'approche.

[1] REDUIT : RÉsumé Dirigé par un Utilisateur Inspiré par une Thématique

Au cours de la construction du résumé, différentes métriques sont utilisées pour définir l'ensemble des unités textuelles les plus représentatives d'un ensemble de documents. La plupart des systèmes définissent une première métrique pour calculer le poids relatif de chaque unité textuelle (phrases, paragraphes …) produite lors de la segmentation des documents. Différents paramètres peuvent être utilisés comme le poids des termes dans les unités textuelles, la catégorie des termes, la longueur des documents … Une seconde fonction permet par la suite de déterminer les unités textuelles pertinentes par rapport à l'ensemble des documents. La pertinence d'une unité textuelle est calculée en fonction de la cooccurrence des termes et de leur degré de couverture par rapport à l'ensemble de la collection. Les différentes unités textuelles jugées pertinentes sont par la suite classées en fonction de leur similarité. Pour cela une mesure de type cosinus est appliquée entre les unités pertinentes pour définir celles qui sont similaires. Enfin une dernière fonction détermine à la fois le degré de redondance de l'information contenue dans les unités textuelles sélectionnées et les unités les plus pertinentes qui vont être retenues pour le résumé (le nombre des unités textuelles sélectionnées dépend de la taille du résumé souhaité par l'utilisateur).

Ce type d'approche a pour avantage d'être relativement simple à développer et peut s'adapter facilement aux différents types de documents. Les résumés produits sont quant à eux bien souvent difficilement exploitables car peu lisibles et manquant de cohérence du fait que le contenu réel des documents n'est pas réellement exploité.

2.2. *Les approches fondées sur l'analyse du contenu des documents*

La seconde catégorie de systèmes dont nous nous sommes inspirés sont les systèmes utilisant des traitements linguistiques sur le contenu des documents et leur structure. Certains systèmes permettent le repérage et la présentation des relations structurales au sein des documents (Olsen *et al.*, 1993 ; Ando *et al.*, 2000), d'autres utilisent des réseaux de neurones (Lin, 1993) ou encore des techniques permettant d'identifier les différentes thématiques contenues dans les documents pour en extraire les passages les plus représentatifs. Néanmoins, quelles que soient les techniques utilisées, tous ces systèmes exploitent le contenu des documents afin d'identifier l'information pertinente qui sera susceptible d'apparaître dans le résumé. Ils effectuent en premier lieu un prétraitement linguistique des documents. Cette étape consiste à effectuer une analyse morphologique et syntaxique des documents. Ainsi sont étiquetés les termes selon leurs catégories grammaticales et sont retenus les mots pleins et les entités nommées. Puis un module d'analyse des composants segmente chaque document en petites unités (phrases) pour en extraire les informations linguistiques les plus pertinentes. La troisième étape consiste à appliquer une fonction de similarité entre les différentes unités pertinentes pour repérer les similarités entre les documents pouvant aller jusqu'à un niveau sémantique. Par exemple, dans le cas où l'ensemble des documents évoque plusieurs thématiques ou décrit différents événements qui évoluent dans le temps, les blocs thématiques parlant d'un même événement seront regroupés ensemble, puis ordonnés chronologiquement afin que le résumé soit le

plus lisible possible pour l'utilisateur. Enfin, un algorithme identifie les expressions les plus représentatives des phrases de chaque thème pour ensuite les placer dans le résumé.

2.3. *Notre positionnement*

Au regard de ces différents travaux, le système REDUIT que nous présentons dans cet article est un système hybride. Il s'appuie en effet à la fois sur une analyse linguistique et thématique des documents et sur un certain nombre de métriques servant de support au processus de filtrage conduisant à la production de résumés multi-documents orientés par des profils utilisateurs.

3. Vue d'ensemble et architecture du système REDUIT

Notre travail (Châar *et al.*, 2002 ; Châar, 2003 ; Châar *et al.*, 2004) met l'accent sur l'utilisation conjointe de profils structurés et d'une analyse discursive des documents pour extraire des passages de textes en rapport avec les attentes de l'utilisateur. La structuration des profils est de nature thématique : un profil est un ensemble de termes répartis entre des thèmes.

L'analyse des documents a pour rôle de mettre en évidence dans les documents à filtrer les structures thématiques pouvant être mises en correspondance avec les thèmes constituant les profils. Elle prend donc la forme d'une analyse thématique permettant de délimiter au sein de chaque document des segments de texte thématiquement homogènes. Elle définit ainsi les unités de base du processus d'extraction qui pourront être appariées avec le profil.

Dans le cadre du processus de filtrage, illustré par la Figure 1, l'analyse des documents est suivie d'une étape d'appariement entre les segments délimités et les thèmes du profil considéré, appariement se fondant sur la similarité de leurs vocabulaires respectifs. Le résultat de cet appariement permet dans un premier temps de définir si un document présente ou non un intérêt du point de vue du profil. Dans l'affirmative, une seconde étape de sélection intervient à un niveau de granularité plus faible. Elle détermine précisément les segments du document s'appariant avec le profil en s'appuyant pour ce faire sur une analyse du rapport global entre le profil et le document. Cette analyse permet en particulier d'étendre le vocabulaire caractérisant le profil à des termes proches figurant dans le document et de sélectionner des segments dont la relation avec le profil est moins directe que la simple identité de termes. On distingue ainsi des segments s'appariant pleinement avec le profil et des segments pour lesquels cet appariement est plus ténu mais qui sont supposés porteurs de davantage d'informations nouvelles en relation avec ce profil.

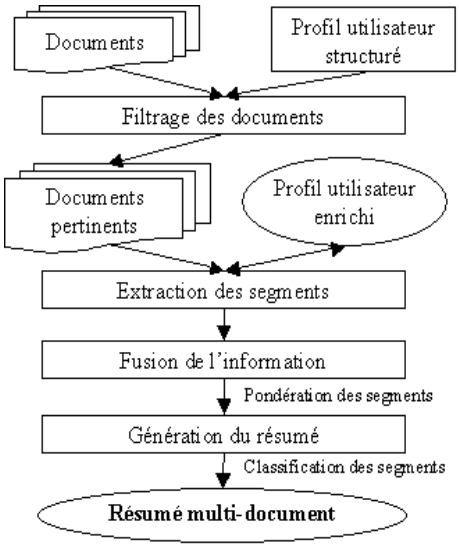

Figure 1. *Architecture du système REDUIT*

Les segments ainsi sélectionnés, qu'ils appartiennent à un même document ou à des documents différents, peuvent être redondants, *i.e.* une même information est susceptible d'apparaître dans plusieurs segments sous des formes légèrement différentes. Une évaluation de la similarité des différents segments sélectionnés est donc réalisée, d'abord au sein d'un même document puis entre documents, en utilisant une mesure de similarité adaptée à la comparaison d'unités textuelles. Pour un ensemble de segments similaires, seul est retenu le segment supposé contenir l'information la plus complète, *i.e.* celui dont le vocabulaire couvre le plus complètement l'ensemble des segments concernés. Cette étape d'élimination des redondances est réalisée en parallèle sur les deux types de segments distingués ci-dessus. En final, le processus de filtrage produit un résumé consistant en une liste de segments pour chacun de ces deux types, segments classés par ordre décroissant d'un score de pertinence par rapport au profil.

4. Les profils utilisateurs

4.1. *Définition de la notion de profil*

Les profils utilisateurs auxquels il est fait référence ici ont pour vocation à représenter les centres d'intérêt des utilisateurs. Contrairement aux requêtes ponctuelles adressées à un moteur de recherche, ils sont utilisés sur une période assez longue, ce qui permet de demander aux utilisateurs un certain investissement

lors de leur constitution et d'aller ainsi au-delà du simple ensemble de mots, cas le plus fréquemment rencontré.

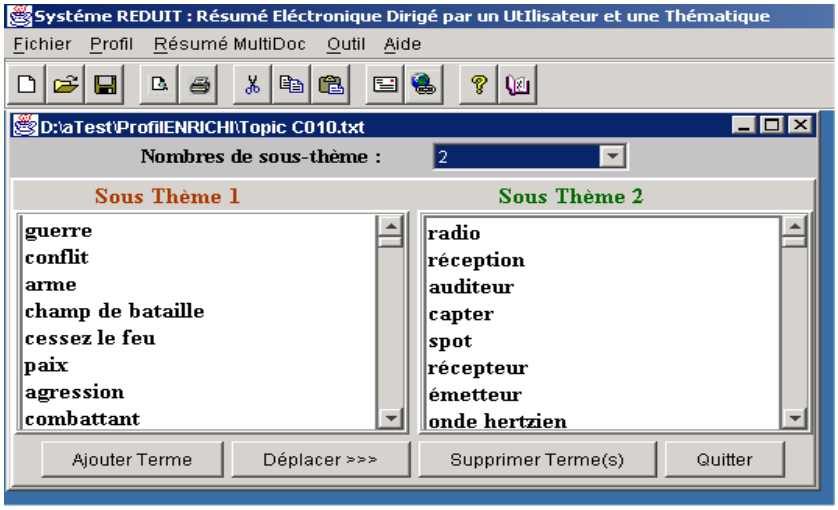

Figure 2. *Exemple de profil sur l'importance et le rôle de la radio en période de guerre*

Nous avons plus précisément choisi d'adopter une structure thématique pour les profils, c'est-à-dire de regrouper les termes qui les composent en sous-ensembles thématiquement homogènes. Cette structuration répond au souci d'améliorer la précision du filtrage en ne mettant pas sur le même plan tous les termes composant un profil. Si l'on s'intéresse au problème du rôle de la radio en période de guerre par exemple (cf. profil de la Figure 2), les documents les plus pertinents seront ceux dans lesquels le vocabulaire lié à la guerre et celui lié à la radio seront simultanément présents. Un document ne comportant que des termes liés à la guerre, même s'ils sont présents en grand nombre, n'a ainsi que peu de chance d'être intéressant. Seule la séparation au niveau du profil des termes liés respectivement à chacun de ces deux thèmes permet d'écarter lors du filtrage un document très marqué par l'un des deux seulement au profil d'un document qui comporte des termes du profil en moins grand nombre mais répartis de façon plus équilibrée entre les deux thèmes qui le caractérisent.

Outre l'amélioration de la précision du filtrage qui en résulte, adopter une telle structuration des profils se justifie par son adéquation avec la nature des demandes de recherche d'information émanant des utilisateurs. Celles-ci sont en effet fréquemment définies par un recoupement de plusieurs thèmes et non par la donnée d'un seul grand thème. Le problème évoqué plus haut du rôle de la radio en période de guerre en est un exemple typique.

Ainsi que le montre la Figure 2, un thème constitutif d'un profil est représenté par un ensemble de termes. Ces termes peuvent être simples ou complexes. Ils sont normalisés au moyen d'une analyse morpho-syntaxique s'appuyant sur les mêmes outils que ceux utilisés pour l'analyse des textes (cf. section 5.1).

4.2. *Structuration thématique des profils*

Même si la structuration des profils en thèmes présente un intérêt du point de vue du résumé automatique comme nous le verrons à la section 9, elle ne correspond pas nécessairement à la façon la plus naturelle pour un utilisateur de spécifier un profil. Ce dernier peut être incité par la forme de l'interface qu'on lui propose à adopter une telle structuration mais le forcer à entrer dans ce schéma pourrait se solder par un rejet global de l'outil, à éviter prioritairement. Le moyen terme le plus raisonnable, lorsque l'utilisateur se contente de fournir une liste non structurée de mots, est de détecter automatiquement la présence éventuelle de plusieurs thèmes au sein de cette liste et de répartir les mots qui la composent selon ces thèmes[2].

Nous avons mis en œuvre cette stratégie en proposant une méthode de classification non supervisée d'un ensemble de mots en thèmes s'appuyant sur un réseau de cooccurrences lexicales. Cette méthode a également été utilisée dans le contexte de l'expansion sémantique pour la recherche d'information (Besançon *et al.*, 2003). Plus précisément, un réseau de cooccurrences lexicales est dans le cas présent un ensemble de cooccurrences lexicales collectées en enregistrant la présence de mots dans une fenêtre de taille fixe que l'on déplace sur un corpus. Le réseau que nous avons utilisé par les expérimentations présentées dans cet article a été construit à partir de 24 mois du journal *Le Monde* sélectionnés entre 1990 et 1994, ce qui représente un corpus d'environ 39 millions de mots. Seules les cooccurrences entre les noms, les verbes et les adjectifs ont été enregistrées. Après filtrage de celles jugées les moins significatives, le réseau obtenu se compose de 23 000 mots et de 5,2 millions de relations. Le lecteur pourra se reporter à (Ferret, 2002) pour de plus amples détails sur ce réseau et sa constitution.

De par ses caractéristiques de construction, un tel réseau est supposé contenir un nombre important de liens de nature thématique, c'est-à-dire rassemblant des mots appartenant à un même thème (cf. (Ferret, 2003) pour une illustration de cette hypothèse). L'idée sous-tendant la méthode de classification proposée est que la densité des liens entre les mots appartenant à un même thème y est plus forte que la densité des liens les reliant à des mots d'autres thèmes. Un sous-thème parmi les mots d'un profil est ainsi identifié par le fait que ses mots forment dans le réseau de cooccurrences lexicales utilisé une sous-composante fortement connexe. Pour

[2] Une rétroaction au niveau de l'utilisateur est aussi envisageable mais en termes d'interface, elle n'apparaît pas très simple si l'on veut offrir à l'utilisateur la possibilité de corriger le « découpage » réalisé automatiquement.

détecter une telle sous-composante, nous nous appuyons sur l'algorithme itératif suivant :

1. sélection des mots du profil à classifier et caractérisation thématique de chacun d'entre eux ;
2. construction d'une matrice de similarité de tous les mots du profil sélectionnés ;
3. construction du sous-thème le plus prégnant ;
4. retour à l'étape 1 en supprimant des mots à classifier ceux formant le dernier sous-thème construit. Le processus s'arrête à ce niveau si le nombre de mots restant pour former un sous-thème est trop faible (*i.e.* inférieur à 3 mots).

La première étape est celle exploitant directement le réseau de cooccurrences lexicales pour associer à chaque mot du profil une représentation thématique constituée des mots du réseau qui lui sont le plus fortement liés dans le contexte de la thématique du profil. Elle vise également à écarter de la classification les mots du profil considérés comme thématiquement non significatifs. Cette étape est, elle aussi, de nature itérative :

1.1. sélection des mots du réseau liés à un nombre minimum de mots du profil, fixé ici à 4 ;

1.2. sélection des mots du profil ayant contribué à la sélection d'un nombre minimum de mots du réseau, fixé ici à 3 ;

1.3. retour à la sous-étape 1.1 en partant des mots du profil issus de la sous-étape 1.2. Le processus s'arrête lorsque l'ensemble des mots sélectionnés, à la fois pour le profil et pour le réseau, est stabilisé.

À l'issue de cette première étape, chaque mot du profil sélectionné se voit associer l'ensemble des mots du réseau qui lui sont liés et qui ont été retenus comme représentatifs de la thématique du profil. Cette représentation sert de base à l'évaluation de la similarité de tous les mots du profil sélectionnés, c'est-à-dire à l'étape 2. La valeur de similarité entre deux mots du profil est donnée par la taille de l'intersection de leurs représentations thématiques. Afin de réduire la sensibilité de la classification au bruit, le vecteur des similarités d'un mot du profil fait l'objet d'un filtrage consistant à ramener à une valeur nulle toutes les valeurs de similarité inférieures à un certain pourcentage, égal ici à 30%, de la valeur maximale des similarités de ce mot.

L'étape 3 se décompose pour sa part en deux phases. La première identifie le mot du profil pouvant servir de graine à un nouveau sous-thème. En l'occurrence, il s'agit du mot dont la somme des valeurs de similarité est la plus forte, c'est-à-dire de celui apparaissant comme le plus central d'un possible sous-thème. La seconde réalise la construction proprement dite du sous-thème en agrégeant à la graine les mots du profil qui lui sont le plus proches. Dans un premier temps, sont rattachés à la graine ceux des mots dont la similarité avec la graine est la plus grande parmi toutes leurs valeurs de similarité non nulles. Pour élargir le sous-thème sans le dénaturer, le

processus est ensuite répété en prenant en compte l'ensemble de ses mots à l'issue du premier rattachement et non plus la seule graine, mais avec une condition restrictive : pour être rattacher au sous-thème, un mot doit aussi avoir une similarité non nulle avec un nombre minimum, ici fixé à 3, de mots déjà rattachés au sous-thème.

<*top*>
<*num*> C013 </*num*>
<*FR-title*> Conférence sur le contrôle des naissances au Caire </*FR-title*>
<*FR-desc*> Quelles discussions et résolutions concernant le contrôle des naissances ont été négociées lors de la conférence sur la population au Caire ? </*FR-desc*>
<*FR-narr*> Toutes les décisions politiques, propositions et résolutions sur le contrôle des naissances de la conférence sur la population sont pertinentes. Les positions des différents pays, organisations et groupes seront également retenues. </*FR-narr*>
</*top*>

sous-thème « conférence »	**sous-thème « contrôle des naissances »**
organisation négocier position proposition conférence résolution	population contrôle naissance

Figure 3. *Un topic CLEF et sa structuration en sous-thèmes*

Dans la perspective plus globale de l'évaluation du système REDUIT (cf. section 9.2), nous avons choisi d'expérimenter l'algorithme de structuration thématique des profils présenté ci-dessus sur la version française des 200 topics[3] élaborés pour les différentes campagnes CLEF d'évaluation en recherche d'information. Pour ce faire, chaque topic a été transformé en une liste de mots pleins en appliquant le même prétraitement linguistique que celui appliqué aux documents (cf. section 5.1). La seule spécificité du traitement appliqué aux topics réside dans l'élimination de certains « méta-mots » (trouver, document, information …) liés à la recherche d'information et non au thème décrit (Besançon *et al.*, 2003). Sur ces 200 topics, l'algorithme a identifié 145 topics ne comportant qu'un seul sous-thème, 48 présentant 2 sous-thèmes et 7 avec 3 sous-

[3] Pour éviter les ambiguïtés liées à une traduction, nous reprendrons le terme consacré « topic » pour désigner le descriptif des requêtes utilisées dans le cadre des évaluations en recherche d'information telles que TREC ou CLEF. Compte tenu du degré de généralité de ces requêtes et du caractère fouillé de leur descriptif, ces topics peuvent être considérés comme des profils.

thèmes. La Figure 3 donne un exemple de la structuration ainsi réalisée pour un de ces topics.

5. L'analyse des documents à filtrer

5.1. *Prétraitement linguistique*

Les documents à filtrer subissent en premier lieu un prétraitement linguistique visant à normaliser leur vocabulaire et faciliter ainsi leur comparaison avec un profil. Cette normalisation consiste à associer à chaque mot d'un document son lemme. Elle est réalisée par des modules de découpage en mots, d'analyse morphologique et d'étiquetage morpho-syntaxique développés au CEA et s'inscrivant dans le prolongement direct de ceux présents dans le système SPIRIT (Fluhr, 1994). Par ailleurs, seuls les mots non grammaticaux susceptibles d'apparaître dans les profils sont sélectionnés, c'est-à-dire les noms, les verbes et les adjectifs. Le système REDUIT s'inscrit sur ce point plutôt dans la perspective des systèmes de recherche d'information que dans celle des systèmes d'extraction d'information : l'identification des thèmes prime sur ce qu'on en dit. En conséquence, REDUIT ne s'attache pas à l'expression de phénomènes tels que les modalités ou les négations par exemple.

Les profils peuvent également contenir des termes complexes, termes d'autant plus importants qu'ils sont généralement plus précis que les termes simples. Pour identifier dans les documents les termes complexes d'un profil, nous nous sommes appuyés sur un ensemble restreint d'heuristiques simples avec l'optique de privilégier le rappel sur la précision. Ces heuristiques sont détaillées à la section 6.1.2. Elles présentent l'avantage de pouvoir être mises en œuvre facilement mais n'offrent évidemment pas la précision d'un outil de reconnaissance de variantes terminologiques tel que FASTR (Jacquemin, 1997) par exemple[4]. Compte tenu des caractéristiques du système REDUIT, nous avons jugé que ce mode de reconnaissance des termes complexes était néanmoins suffisant. Une évaluation précise de l'intérêt d'un outil de type FASTR dans le contexte du filtrage resterait toutefois à réaliser.

[4] FASTR permet d'identifier des variantes terminologiques échappant aux heuristiques que nous avons mises en place et pourrait donc encore améliorer le rappel de la reconnaissance des termes des profils. Nous avons cependant montré, dans le contexte d'un système de question/réponse pour l'anglais (de Chalendar *et al.*, 2003), que ce type d'outil est également plus sensible aux erreurs d'étiquetage morpho-syntaxique qu'une approche à base d'heuristiques, ce qui est source d'un certain silence.

5.2. *Analyse thématique*

L'analyse thématique des documents a dans le cas présent pour rôle de segmenter les documents en unités thématiquement homogènes, tâche que l'on nomme segmentation thématique. Les segments ainsi délimités constituent les unités textuelles de base qui sont comparées aux profils et le résultat du filtrage prend la forme d'une liste restreinte de ces segments, ordonnés suivant leur pertinence par rapport au profil considéré.

Plusieurs systèmes de segmentation thématique ont été développés ces dernières années (Hearst, 1997 ; Kan *et al.*, 1998 ; Choi, 2001 ; Utiyama *et al.*, 2001) mais les implémentations disponibles ont été réalisées pour l'anglais en s'appuyant essentiellement sur un processus de racinisation pour normaliser le vocabulaire des textes, technique peu adaptée à des langues telles que le français du fait de leur morphologie plus riche. Nous avons donc choisi de réimplémenter le système C99 de Choi (Choi, 2000) en l'adaptant afin de prendre en entrée le résultat du prétraitement des textes décrit à la section précédente. Le choix de C99 s'appuie sur le bon rapport complexité / efficacité de son algorithme. Dans une première phase, une matrice de cohésion des phrases du texte considéré est construite : chaque phrase est transformée en un vecteur contenant les mots sélectionnés à l'issue du prétraitement linguistique et sa cohésion avec chacune des autres phrases du texte est évaluée par une mesure vectorielle. Après application d'un filtre destiné à renforcer le contraste des valeurs, la segmentation proprement dite est réalisée par une classification divisive des phrases s'appuyant sur la matrice de cohésion : à chaque itération, une nouvelle borne délimitant deux segments est posée à l'endroit maximisant la cohésion interne de ces deux segments. Le processus se poursuit tant que la pose de nouvelles bornes s'accompagne d'une augmentation de la cohésion globale. Les segments obtenus ont une taille moyenne aux alentours d'une centaine de mots, cette valeur variant bien entendu en fonction des variations thématiques effectivement rencontrées dans les textes.

Nous n'avons abordé dans cette partie que l'analyse thématique réalisée en dehors du contexte d'un profil particulier. Les opérations d'identification thématique[5] que nous décrirons dans la section 6 permettant d'établir qu'un segment s'apparie avec un profil ou plus globalement que le thème principal d'un texte s'apparie avec un profil relèvent également de l'analyse thématique. Cependant, elles ne sont pas réalisées ici selon une perspective générique qui consisterait en premier à caractériser le thème d'un segment ou d'un texte par ses mots les plus caractéristiques pour ensuite comparer ceux-ci avec les mots du profil.

[5] L'identification thématique est la partie de l'analyse thématique visant à déterminer le thème d'une unité textuelle.

6. Filtrage des documents en fonction d'un profil utilisateur

À la suite du prétraitement linguistique et de l'analyse thématique, les documents font l'objet du filtrage proprement dit dont le but est de sélectionner les segments les plus intéressants par rapport au profil considéré. Ce filtrage s'opère en deux temps : une première étape permet d'écarter les documents sans rapport avec le profil ; la seconde assure la sélection de l'ensemble des segments en relation avec le profil.

6.1. *Sélection des documents*

6.1.1. *Principes*

Nous considérons que trois grands cas de figure sont à distinguer en termes de filtrage d'un document :

– le document est globalement en relation avec le profil, même s'il peut comporter des parties abordant des thèmes en dehors de ceux du profil ;

– une partie seulement du document est en relation avec le profil. Ce dernier correspond à un thème secondaire du document et n'est donc évoqué que ponctuellement au sein de celui-ci ;

– le document n'a pas de relation avec le profil, même à un niveau local.

L'étape de sélection des documents vise à séparer les documents relevant des deux premiers cas de figure de ceux relevant du dernier. Nous considérons par ailleurs que les documents globalement en relation avec le profil ont nécessairement une partie s'appariant avec ce profil (cf. section 0). Par conséquent, le critère de sélection des documents que nous appliquons est lié au deuxième cas de figure : un document est sélectionné à condition que l'un au moins de ses segments (issus de l'analyse thématique) s'apparie avec le profil considéré.

6.1.2. *Appariement d'un profil et d'un segment*

Ainsi que nous l'avons indiqué à la section 4, la structuration thématique des profils vise à éviter qu'un document ou une partie de document ne soit sélectionnée alors qu'elle n'aborde qu'une des dimensions d'un profil. Il apparaît donc logique d'imposer qu'un segment de texte ne puisse s'apparier avec un profil que si chacun des thèmes constituant ce profil est représenté dans le segment. Les segments étant en moyenne de petite taille (cf. section 5.2), nous considérons que la représentation d'un thème dans un segment se manifeste par la présence au sein de celui-ci d'un terme caractérisant ce thème au niveau du profil. Ce critère peut apparaître au premier abord un peu faible à l'échelle d'un seul thème mais il est beaucoup plus significatif à l'échelle d'un profil regroupant plusieurs thèmes. Bien que les termes complexes soient en général plus informatifs que les termes simples, nous n'imposons pas que le terme représentant un thème dans un segment soit un terme

complexe car nous ne voulons pas imposer de contrainte trop stricte sur le contenu des thèmes définis par l'utilisateur du système de filtrage. Par ailleurs, les termes complexes posent un problème spécifique de reconnaissance que nous avons choisi de résoudre en adoptant un compromis moyen entre précision et rappel (cf. section 5.1). La reconnaissance stricte d'un terme complexe *TC* du profil dans un segment obéit plus précisément à la série d'heuristiques suivantes :

– les termes simples TS_i composant *TC* doivent apparaître dans le segment dans le même ordre qu'au sein de *TC*[6]. La reconnaissance des TS_i est directe puisqu'ils sont normalisés de la même façon dans les documents et dans le profil ;

– soit *N*, le nombre des TS_i. L'espace occupé par une occurrence de *TC* dans le segment ne peut dépasser 1,5 * *N* mots pleins. Ce facteur permet de prendre en compte d'éventuelles variations syntaxiques de type insertion ;

– si *TC* comporte des prépositions, celles-ci doivent être également présentes au niveau de ses occurrences dans les documents, ceci en respectant leur position par rapport aux TS_i. Par ailleurs, une occurrence présumée de *TC* ne doit contenir aucun signe de ponctuation.

Le terme complexe *TC* peut également être reconnu lorsque seulement un de ses sous-termes *ST* est présent. On parle alors de reconnaissance approchée. Trois conditions doivent être remplies pour ce faire :

– *ST* doit regrouper au moins 50% des termes simples de *TC* ;

– une occurrence de *ST* est reconnue si elle respecte les trois contraintes de reconnaissance stricte d'un terme ;

– *TC* doit avoir été reconnu de façon stricte au moins une fois dans le document auquel appartient le segment considéré.

Dans le cadre de l'identification d'un thème dans un segment, la reconnaissance d'un terme caractéristique de ce thème peut être stricte ou bien approchée.

6.2. *Sélection d'extraits de documents*

6.2.1. *Principes de la sélection des segments*

Le fait qu'un document a été sélectionné renvoie à deux cas de figure comme nous l'avons vu à la section 6.1.1 : le document s'apparie globalement avec le profil considéré ou bien seulement une partie de ce document s'apparie avec ce profil. Dans le second cas, la thématique du profil n'étant évoquée que secondairement dans le document, il n'y a pas lieu de chercher à sélectionner d'autres segments que

[6] Ce type d'heuristique est bien évidemment plus ou moins efficace selon la langue : l'ordre des mots au sein d'un mot composé semble ainsi plus stable d'une forme à une autre en français qu'en anglais.

celui ou ceux s'appariant avec le profil selon les critères énoncés à la section 6.1.2. Dans le premier cas au contraire, il peut être intéressant d'élargir le champ de la sélection à des segments ne répondant pas strictement aux critères d'appariement avec le profil mais contenant des termes du document jugés liés à ceux du profil. C'est en particulier une façon de s'ouvrir vers la détection de nouvelles tendances en relation avec les thèmes du profil.

Plus précisément, lorsqu'un document s'apparie avec un profil, un de ses segments est sélectionné dès lors qu'il contient un terme représentant chacun des thèmes de ce profil, ce terme pouvant faire partie de la description du profil spécifiée par un utilisateur ou bien être un terme inféré, c'est-à-dire un terme du document considéré comme lié aux termes du profil décrivant le thème en question.

6.2.2. *Sélection des termes inférés*

La liaison entre un terme inféré et un terme du profil s'appuie sur une série de cooccurrences au sein du document. Plus précisément, soit $\{tp_{Ki}\}$, l'ensemble des termes du thème K appartenant au profil qui sont présents dans le document. On définit l'ensemble $\{td_{Ki}\}$ des termes du document tels que td_{Ki} cooccurre avec un terme tp_{Ki} (pas nécessairement le même d'un segment à l'autre) dans un segment et ce, dans une proportion suffisamment importante de segments du document. Cette proportion a été fixée dans le cas présent à 1/3. Les termes td_{Ki} constituent ce que nous avons appelé ci-dessus des termes inférés.

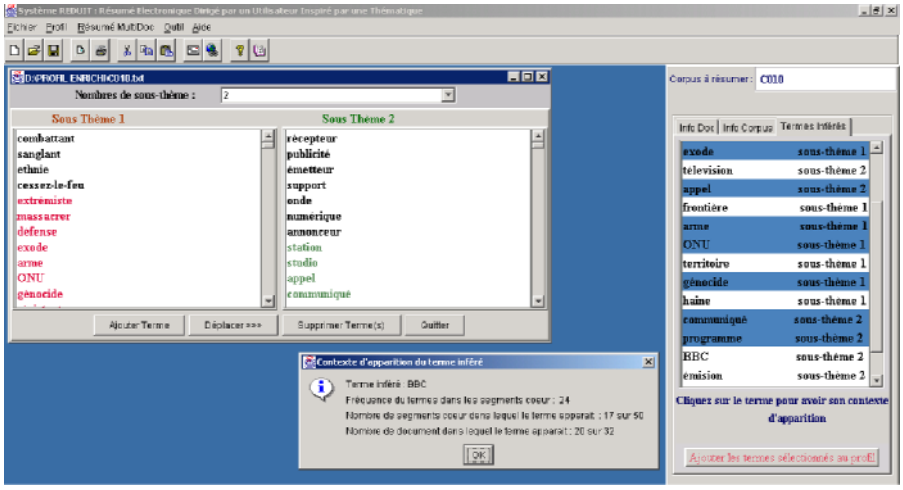

Figure 4. *Exemple de termes inférés pour le profil de la Figure 2 sur le rôle de la radio en période de guerre*

Ces termes inférés représentent une forme d'adaptation des profils par rapport aux documents auxquels ils sont confrontés dans le cadre du processus de filtrage. Les profils ont tendance à donner une description assez générale des thèmes qu'ils recouvrent. Au sein de chaque document, on retrouve les termes de cette description mais on y trouve également une caractérisation beaucoup plus spécifique qu'il est nécessaire de prendre en compte pour analyser finement l'organisation thématique du document. La mise en évidence de ces termes inférés possède à cet égard une certaine parenté avec le processus de blind relevance feedback utilisé en recherche documentaire.

La Figure 4 donne une illustration des termes inférés extraits de documents, en l'occurrence ceux du corpus CLEF 2003 (cf. section 9), pour le profil de la Figure 2 sur le rôle de la radio en période de guerre. On voit ainsi apparaître (termes en couleur dans les sous-thèmes ou termes dans la fenêtre de droite) des termes comme *extrémiste*, *massacrer*, *arme* … liés au sous-thème de la guerre et des termes tels que *station*, *studio*, *communiqué* … concernant plus spécifiquement le sous-thème de la radio. En pratique, l'utilisateur a la possibilité de valider ou d'écarter les termes inférés extraits automatiquement afin d'améliorer la qualité du filtrage.

6.2.3. *Appariement d'un profil et d'un document*

Dire qu'un document s'apparie globalement avec un profil revient à dire que le thème principal du document correspond au thème représenté par ce profil[7]. Bien que le problème de la formalisation de l'organisation thématique des textes ait été peu exploré, les travaux sur le résumé automatique laissent entrevoir une définition opérationnelle de la notion de thème principal que l'on peut énoncer comme suit : le thème principal d'un texte est le thème abordé en début ou/et en fin de texte et qui est l'objet d'une partie importante de celui-ci.

En transposant cette définition dans notre contexte, déterminer si le thème principal d'un document s'apparie avec un profil revient à vérifier les deux conditions suivantes :

– le profil doit s'apparier avec le premier et/ou le dernier segment du document ;

– plus globalement, l'ensemble des segments du document s'appariant avec le profil doit représenter une part significative de l'ensemble des segments du document.

La première condition fait appel à la notion d'appariement entre profil et segment développée à la section 6.1.2. La seconde s'appuie quant à elle sur la version élargie de cet appariement exposée à la section 6.2.1, version prenant en compte les termes inférés, c'est-à-dire des termes du document liés aux termes du profil. En outre, la part significative imposée par cette seconde condition a été fixée à 1/3 des segments du document.

[7] Le thème représenté par un profil correspond à la réunion des thèmes qui le composent, ce qui présuppose une certaine compositionnalité des thèmes.

7. Fusion des informations

L'objectif global du filtrage d'information est non seulement de sélectionner des extraits de document en relation avec un profil mais également de maximiser le nombre d'informations dont un opérateur peut prendre connaissance en lisant une certaine quantité de texte. L'étape de sélection des segments de texte détaillée dans la section 6 doit donc être suivie d'une étape de fusion visant à minimiser les redondances entre segments. Cette fusion, qui intervient d'abord entre les segments d'un même document puis entre les segments provenant de plusieurs documents, est réalisée par le choix des segments les plus représentatifs des informations véhiculées.

7.1. *Fusion intra-document*

Du point de vue de la détection des redondances entre segments, nous distinguons deux classes de segments :

– les segments dont l'appariement avec le profil se fonde sur des termes du profil reconnus de façon stricte dans le segment ;

– les segments dont l'appariement avec le profil se fonde au moins partiellement sur des termes inférés.

Les premiers, appelés *segments cœur*, représentent plutôt l'instanciation dans un document particulier de l'information déjà contenue dans le profil tandis que les seconds, appelés *segments extension*, ont davantage vocation à apporter de l'information nouvelle en relation avec le profil. Ces deux classes étant complémentaires, nous ne chercherons pas à fusionner les informations issues d'une classe avec celles provenant de l'autre classe. Au sein de chacune de ces deux classes, la détection des redondances entre segments s'appuie sur une mesure de similarité à laquelle est associé le seuil fixé *a priori* S_{fusion}. Nous avons classiquement choisi la mesure du cosinus, qui s'avère particulièrement bien adaptée à la comparaison d'unités textuelles. La similarité entre deux segments S_1 et S_2 est donc donnée par :

$$sim(S_1, S_2) = \frac{\sum_i nbOcc(t_i, S_1) \cdot nbOcc(t_i, S_2)}{\sqrt{\sum_i nbOcc(t_i S_1)^2 \cdot \sum_i nbOcc(t_i, S_2)^2}} \quad [1]$$

avec $nbOcc(t_i, S_{\{1,2\}})$, le nombre d'occurrences du terme t_i dans le segment $S_{\{1,2\}}$. Les termes t_i considérés ici correspondent aux lemmes issus du prétraitement linguistique des documents. Si cette mesure de similarité dépasse le seuil S_{fusion}, les deux segments sont jugés similaires et sont donc supposés contenir en première

approximation les mêmes informations. Un seul des deux segments peut donc représenter les deux. Dans le cas contraire, on conserve les deux segments.

Plus globalement, la mesure de similarité [1] est évaluée entre tous les segments au sein de chacune des deux classes distinguées ci-dessus (segments cœur et segments extension). Ces segments sont ensuite regroupés en fonction de la valeur de cette mesure : chaque segment est associé au segment avec lequel sa similarité est la plus forte, à condition toutefois que cette similarité soit supérieure au seuil S_{fusion}. On obtient ainsi un ensemble de regroupements disjoints de segments similaires. Dans le cas où la similarité entre tous les segments est assez forte, une classe peut ne comporter qu'un seul groupe de segments.

Un représentant est ensuite sélectionné pour chaque groupe de segments ainsi défini. Ce segment doit être le plus représentatif possible des informations véhiculées par les segments du groupe. Nous nous appuyons pour ce faire sur le vocabulaire caractérisant le groupe, vocabulaire défini comme l'ensemble des termes simples communs à au moins deux de ses segments. Le représentant du groupe est plus précisément le segment abritant la plus large proportion de ce vocabulaire.

7.2. *Fusion inter-document*

À l'issue de l'étape de fusion intra-document, chaque document est représenté par deux ensembles rassemblant des segments non similaires selon [1]. La fusion inter-document commence par l'union des ensembles de même type des différents documents considérés. Au sein de chacun de ces deux ensembles, nous appliquons le même algorithme de regroupement et de sélection d'un représentant que celui appliqué dans le cadre de la fusion intra-document. En final, nous obtenons donc un ensemble de segments cœur et un ensemble de segments extension.

7.3. *Ordonnancement des extraits de documents*

Afin de faciliter l'exploitation ultérieure des deux ensembles de segments issus du filtrage, en particulier leur visualisation, un ordre de pertinence est défini sur leurs éléments. En revanche, ces deux ensembles sont conservés disjoints dans la mesure où chacun d'eux rend compte d'une dimension particulière du filtrage, dimension qui sera ou non exploitée suivant l'application dans laquelle celui-ci vient s'insérer.

Plus précisément, chaque segment se voit attribuer un score calculé sur la base du vocabulaire qui le compose. Ce score prend en compte à la fois la présence des termes du profil, sous une forme stricte ou approchée, ainsi que celle des termes communs au groupe de segments dont celui considéré est le représentant. Ce score est donné par :

$$score(S) = 1{,}0 \cdot \sum_i nbOcc(tps_i, S) + 0{,}75 \cdot \sum_i nbOcc(tpa_i, S) + 0{,}5 \cdot \sum_i nbOcc(tcg_i, S) \quad [2]$$

où tps_i est un terme du profil reconnu de façon stricte, tpa_i est un terme du profil reconnu de façon approchée et tcg_i est un des termes communs aux segments formant le groupe dont le segment considéré est le représentant. Les pondérations adoptées favorisent la proximité par rapport au profil tout en accordant une place significative aux termes qu'il semble raisonnable d'associer au profil (termes tcg_i). Les segments ayant des tailles assez similaires, nous n'avons pas adopté de normalisation en fonction de la taille des segments.

Finalement, le résultat du processus global de filtrage se présente sous la forme de deux listes, celle des segments cœur et celle des segments extension, ordonnées suivant l'ordre décroissant du score de leurs segments. Il est à noter que le score défini par [2] sert également de support à l'application éventuelle d'un taux de compression par l'utilisateur. De par son fonctionnement, REDUIT détermine de lui-même le nombre de segments qu'il présente à l'utilisateur. Néanmoins, si ce dernier veut imposer un taux de compression supérieur, ne lui seront alors présentés que les segments de plus fort score dans la limite du taux de compression fixé.

8. Construction et présentation du résumé[8]

La construction et la présentation du résumé final fourni par notre système prennent en compte certains critères dits de bonne formation ainsi que les contraintes issues du contexte applicatif considéré. Le système REDUIT s'inscrivant dans le cadre plus général d'un système de veille technologique, nous nous sommes essentiellement attachés à la façon de regrouper l'information sélectionnée en fonction de son utilité du point de vue de la veille : les segments cœur représentent l'information de fond en relation directe avec l'objet de la veille tandis que les segments extension sont le lieu de découverte de nouvelles tendances, validées par la présence simultanée des segments cœur. Au sein de chaque type d'information, le critère principal de tri est celui de la pertinence par rapport au profil mais dans une plage de pertinence donnée, la construction du résumé final prend également comme critères d'ordonnancement secondaires la position des segments dans les documents et la date de parution des documents. En revanche, nous n'avons pas intégré de contraintes liées à la cohésion du résumé produit comme le fait par exemple de tenir

[8] Nous utilisons le terme « résumé » pour désigner le résultat final de notre système de filtrage tout en étant conscient qu'il s'agit partiellement d'un abus de langage puisque ce « résumé » est composé d'une suite de segments textuels et n'a pas vocation à constituer un texte cohérent.

compte des anaphores. Le fait de travailler avec des segments rend en effet ce dernier point moins important que dans le cas d'unités de taille plus petite.

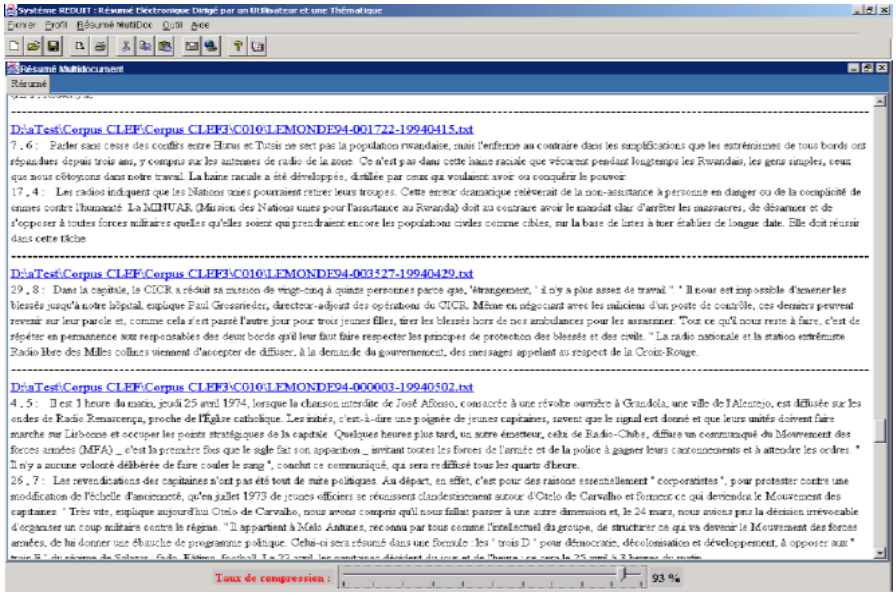

Figure 5. *Exemple de résumé produit par le système REDUIT*

Par ailleurs, les interfaces développées jouent un rôle primordial pour la facilité d'appréhension du résumé final. Elles permettent à l'utilisateur d'accéder facilement aux passages constituant dans le résumé, de « zoomer » sur ces derniers et de revenir au texte source afin de restituer rapidement le contexte des segments sélectionnés et juger ainsi d'éventuelles incohérences avec le profil. L'utilisateur peut également contrôler la taille du résumé produit, taille déterminant le nombre de segments visualisés pour chaque regroupement de segments. Enfin, les interfaces offrent la possibilité de rétroagir sur le résultat en confirmant ou non la pertinence des segments affichés, ce qui permet en particulier de valider indirectement les termes inférés sélectionnés.

9. Expérimentation et discussion

9.1. *Modalités de l'évaluation*

Ainsi que le rappelaient Radev, Hovy et McKeown en introduction d'un numéro spécial de la revue Computational Linguistics sur le résumé automatique (Radev *et al.*, 2002), évaluer un résumé est une tâche difficile dans la mesure où la notion de

résumé idéal n'existe pas. De ce fait, même des évaluateurs humains ne sont en accord les uns avec les autres que dans 60% des cas lorsqu'il s'agit de juger du recouvrement du contenu de phrases issues d'articles de journaux. Les travaux sur l'évaluation du résumé automatique font généralement la distinction entre les évaluations intrinsèques et les évaluations extrinsèques (Sparck Jones *et al.*, 1995). Les premières ont pour objectif de mesurer la qualité des résumés produits, en général en les comparant à des résumés de référence élaborés par des humains. Les secondes mesurent dans quelle mesure les résumés produits peuvent contribuer à la réalisation d'une tâche et l'améliorer. Par ailleurs, les évaluations intrinsèques peuvent se focaliser sur la forme même des résumés (cohérence globale, reprises anaphoriques …) ou sur leur contenu.

Pour l'évaluation du système REDUIT, nous avons adopté une approche intrinsèque s'appuyant sur le contenu. Plus précisément, nous avons été amenés à mettre en place un protocole spécifique mais s'inspirant des évaluations de même type déjà réalisées dans le domaine du résumé automatique. Les évaluations existantes, SUMMAC (Mani *et al.*, 1998) et DUC (2001, 2002 et 2003) pour l'anglais ou encore la tâche résumé des workshops NTCIR 2 et 3 pour le japonais (Fukusima *et al.*, 2001), ne se sont en effet pas intéressées au français et n'ont par ailleurs pas proposé de tâche conjuguant les différentes caractéristiques du système REDUIT : production de résumé multi-documents guidés par un profil ou une requête avec un niveau de granularité de l'ordre du passage.

Topic 10	**Influence des messages radio** en *période de guerre*
Topic 31	La **protection juridique** des *consommateurs européens*
Topic 62	**Conséquence naturelle** des *tremblements de terre* au JAPON
Topic 159	**Influence sur l'économie** de *la pollution* des SITES PETROLIERS

Tableau 1. *Exemples de topics (**en gras** le 1er sous-thème, en italique le 2ème sous-thème, et en MAJUSCULE le 3ème sous-thème) retenus pour l'évaluation de REDUIT*

Pour notre évaluation, les profils sont constitués, comme dans le cas de la tâche *Adhoc* de SUMMAC, par des topics tels que ceux utilisés dans les tâches de recherche d'information. Les topics TREC utilisés par SUMMAC ont été remplacés dans notre cas par des topics CLEF. Ceux-ci présentent l'avantage pour nous d'exister pour le français et plus généralement d'être déclinés dans un nombre important de langues. Pour l'évaluation du système REDUIT, nous avons retenu un ensemble de 14 topics ayant comme spécificité de regrouper deux et trois thématiques différentes. Notre objectif est en effet d'illustrer l'intérêt de la prise en compte de cette pluralité thématique lorsqu'elle existe. Le Tableau 1 donne quelques exemples de ces topics au travers de leur intitulé.

Pour chaque topic, nous disposons grâce à l'évaluation CLEF d'un ensemble de documents validés comme pertinents par rapport au topic et d'un ensemble de documents validés comme non pertinents. Ces documents sont des articles du journal *Le Monde* des années 1994 et 1995 et des dépêches de l'agence SDA couvrant la même période. Chaque document pertinent d'un topic (une vingtaine en moyenne par topic) a fait l'objet d'une annotation automatique afin d'en délimiter les unités textuelles élémentaires du point de vue du résumé, à l'image de ce qui est fait dans DUC. Nous avons en pratique retenu la phrase comme unité élémentaire. Un exemple de ce balisage (balises <SU>) est donné par la Figure 6. Ce premier balisage a été complété par une annotation manuelle des unités textuelles considérées comme pertinentes par rapport au topic (balises <PER>). Le résumé de référence pour chaque document d'un topic est donc constitué de la suite des unités textuelles, *i.e.* SUs, sélectionnées comme pertinentes par rapport à ce topic.

> […] <SU id="11"> Les témoignages des personnes présentes sur le terrain sont désormais appuyés par les Nations Unies . </SU><SU id="12"> Rapporteur spécial de la commission des droits de l' homme à l' ONU , M. Dégni-Ségui consacre un article de son rapport sur la situation des droits de l' homme au Rwanda , daté du 12 août , à l' " action de la RTLM " : </SU>[…] **<PER>** <SU id="15"> Mensonges , propagande , désinformation ... </SU><SU id="16"> Les statuts de la première radio libre rwandaise , signés le 30 septembre 1993 par le ministre de l' information , Faustin Rucogoza , et par Félicien Kabuga , président du comité d' initiative de la RTLM . </SU><SU id="17"> Dans l' alinéa 2 de l' article 5 , la radio s' engageait " à ne pas diffuser les émissions de nature à inciter à la haine , à la violence et à toute forme de division " ... </SU> **</PER>**[…] **<PER>** <SU id="18"> Mais la radio est devenue une arme , comme l' argent et l' armée , dont les responsables de diverses factions en lutte pour le pouvoir se disputent le monopole (le Monde du 21 juillet) . </SU><SU id="19"> Ainsi , la seule radio rwandaise autorisée _ Radio Mohabura _ est " héritée des structures du FPR " , concède son directeur des programmes , un ancien militaire du Front patriotique . </SU><SU id="20"> La guerre des ondes </SU> **</PER>** […]

Figure 6. *Passage d'un document de référence pour l'évaluation de REDUIT (topic 10)*

Au fur et à mesure du développement des évaluations relatives au résumé automatique, un certain nombre de métriques ont été proposées allant des mesures de précision et de rappel classiquement utilisées en recherche d'information à des mesures plus spécifiques telles que la mesure d'utilité relative, proposée dans (Radev *et al.*, 2000), ou plus récemment la mesure ROUGE (Lin *et al.*, 2002 ; Lin *et al.*, 2003), développée à l'occasion des dernières évaluations DUC. Pour l'évaluation de REDUIT, nous avons retenu les mesures de précision et de rappel et non une mesure

plus spécifiquement dédiée au résumé car ces dernières sont davantage adaptées à la production de résumés courts qu'à l'extraction de passages. Nous avons adopté en cela la même perspective que (Alonso *et al.*, 2002).

Dans le contexte qui est le nôtre, la précision et le rappel sont définis par les rapports :

$$\begin{aligned} \text{Précision} &= P/(NP + P) \\ \text{Rappel} &= P/(P + R) \end{aligned} \quad [3]$$

où *NP* est le nombre d'unités non pertinentes fournies par le système, *P,* le nombre d'unités pertinentes fournies par le système et *R*, le nombre d'unités pertinentes dans le corpus de référence et non fournies par le système. Dans le cas de l'évaluation du filtrage (cf. section 9.2.1), les unités sont des documents ; dans le cas de l'évaluation de l'extraction de segments (cf. section 9.2.2), il s'agit de SUs.

$$\text{F1} - \text{mesure} = \frac{2*\text{Précision}*\text{Rappel}}{\text{Précision} + \text{Rappel}} \quad [4]$$

La f1-mesure, moyenne harmonique du rappel et de la précision, est utilisée de manière classique pour synthétiser ces deux mesures en un seul indicateur. Les résultats présentés dans la section suivante correspondent aux moyennes des résultats obtenus pour les 14 topics de test.

9.2. *Résultats de l'évaluation*

Nous avons évalué séparément les deux grandes fonctions du système REDUIT, à savoir le filtrage de documents (cf. section 6.1) et l'extraction de passages par rapport à un profil (cf. section 6.2). Dans les deux cas, nous avons cherché à mettre en évidence l'intérêt de la prise en compte de l'hétérogénéité thématique des profils et des textes en comparant les résultats obtenus avec des profils non structurés sur le plan thématique et des profils dans laquelle cette structuration est mise en évidence. Pour rendre cette comparaison la plus objective possible, la structuration des 14 profils retenus pour l'évaluation a été réalisée de manière automatique selon la méthode exposée à la section 4.2.

9.2.1. *Évaluation de la méthode de filtrage des documents*

L'objectif de cette évaluation est d'illustrer la capacité du système REDUIT à sélectionner des documents en relation avec un profil. Le corpus de référence que nous utilisons est constitué des 3780 documents pour lesquels nous disposons, grâce au résultat des évaluations CLEF, d'un jugement par rapport aux 14 topics retenus pour l'évaluation. Il est à noter que ce corpus peut être considéré comme

particulièrement difficile dans la mesure où cet ensemble rassemble, conformément à la technique du pooling utilisé dans les évaluations de type TREC, les documents considérés comme les plus pertinents par les systèmes de recherche d'information ayant participé à CLEF. Sur ces 3780 documents, seuls 320 sont véritablement pertinents vis-à-vis des 14 topics sélectionnés.

Méthode de filtrage	Rappel	Précision	F1-mesure
REDUIT (v0)	0,89	0,11	0,21
REDUIT (v1)	0,82	0,44	0,57

Tableau 2. *Résultats de l'évaluation du processus de filtrage des documents*

Le Tableau 2 donne les résultats obtenus avec le système REDUIT lorsque chaque profil est considéré comme ne comportant qu'un seul thème (v0) et lorsque la structuration thématique exposée à la section 4.2 est appliquée. Comme on pouvait s'y attendre, la prise en compte de l'hétérogénéité thématique des profils se caractérise par une amélioration significative de la précision, laquelle se fait au détriment d'une légère baisse du rappel. Le gain global est néanmoins net. Par ailleurs, la faiblesse des valeurs de précision n'est pas surprenante : outre le caractère « difficile » du corpus utilisé, le filtrage de documents opéré à partir d'un profil défini manuellement ou par l'intermédiaire d'un topic de style TREC est reconnu comme étant une tâche difficile, au point d'avoir été abandonnée dans l'évaluation Filtering de TREC (Hull *et al.*, 2000).

9.2.2. *Évaluation de la sélection des unités textuelles pour la construction des résumés*

Ce second volet de l'évaluation a pour objet de mesurer la capacité du système REDUIT à sélectionner les parties d'un document les plus directement liées à un profil. Celles-ci sont constituées par les segments définis par l'analyse thématique décrite à la section 5.2 que REDUIT retient au niveau de chaque document. Compte tenu de la granularité de la segmentation thématique et de l'indépendance de la référence manuelle par rapport à cette segmentation, nous avons ajouté une heuristique supplémentaire : pour chaque segment sélectionné, REDUIT ne retient que ceux des SUs qui le composent comportant au moins un mot du profil. Comme précédemment, REDUIT (v0) est le système faisant abstraction de la structure thématique des profils, au contraire de REDUIT (v1). Les résumés obtenus avec REDUIT (v0) sont donc constitués par l'ensemble des segments thématiques contenant des termes du profil, sans restriction quant à la représentation de l'ensemble de ses thématiques. Il est à noter que pour que la comparaison de REDUIT (v0) et REDUIT (v1) ne soit pas biaisée, les conditions portant sur le

nombre de termes du profil conditionnant la sélection d'un segment sont les mêmes dans les deux cas.

Pour faciliter la mise en perspective des résultats du Tableau 3 concernant le système REDUIT, les résultats de trois méthodes pouvant être considérées comme des références basses sont également fournis :

– *référence basse 1* se contente de sélectionner toujours le premier segment de chaque document ;

–. *référence basse 2* sélectionne toujours le premier et le dernier segment de chaque document ;

– .*référence basse 3* sélectionne toujours le dernier segment de chaque document.

Ces méthodes, adaptées de celles définies dans DUC, s'appuient sur la constatation que l'introduction et la conclusion d'un texte constituent souvent une forme de résumé ou tout du moins, rassemblent une grande part des informations importantes contenues dans le texte.

Méthodes d'extraction	**Rappel**	**Précision**	**F1-mesure**
référence basse 1	0,56	0,36	0,44
référence basse 2	0,68	0,34	0,45
référence basse 3	0,11	0,23	0,14
REDUIT (v0)	0,68	0,53	0,60
REDUIT (v1)	0,67	0,65	0,65
REDUIT (v2)	0,82	0,60	0,70

Tableau 3. *Résultats de l'évaluation de l'extraction de segments*

Comme dans le cas du filtrage de document, le Tableau 3 montre l'intérêt de la prise en compte d'une structuration thématique des profils, avec un impact similaire sur les résultats : la précision augmente de façon importante tandis que le rappel baisse légèrement, ce qui se traduit globalement par une nette amélioration. On constate également que toutes les versions de REDUIT dépassent assez largement les méthodes de référence basse, même si le rappel de *référence basse 2* est comparable à celui de REDUIT (v0) et de REDUIT (v1). Ce dernier point justifie d'ailleurs *a posteriori* les conditions posées à la section 6.2.3 concernant l'appariement global d'un profil et d'un document. La dernière ligne du Tableau 3, REDUIT (v2), correspond quant à elle à la version complète du système REDUIT, c'est-à-dire une version intégrant la prise en compte des mots inférés. Étant donné que nous n'avons pas pour le moment défini une manière automatique de rattacher

un mot inféré à un sous-thème d'un profil[9], cette affectation a été faite manuellement pour cette évaluation. Les résultats de REDUIT (v2), en nette amélioration par rapport à ceux de REDUIT (v1) en termes de rappel et stables en termes de précision, montrent que l'intégration des mots inférés dans les profils et leur prise en compte spécifique dans le processus d'extraction des segments est d'un intérêt potentiel réel.

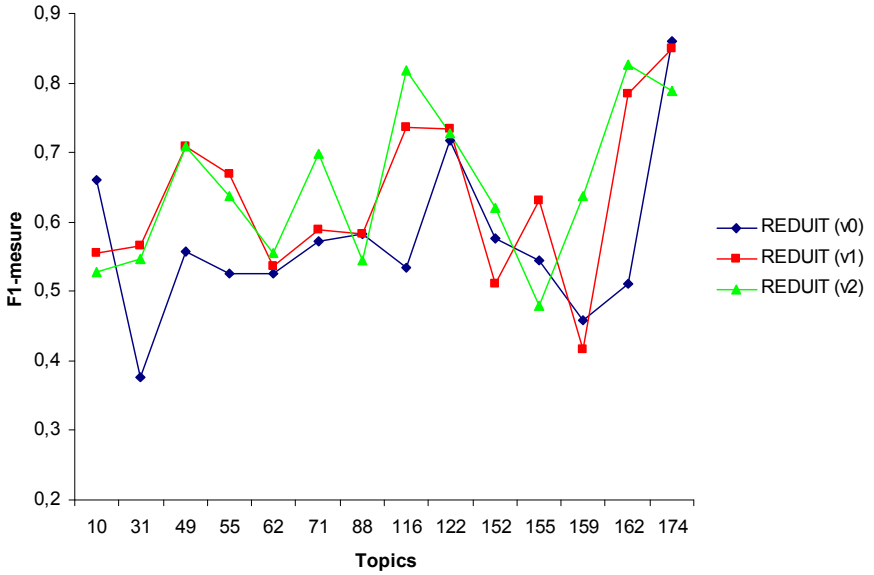

Figure 7. *Détail au niveau de chaque topic de la f1-mesure pour les 3 versions de REDUIT*

La Figure 7 montre enfin les différences inter-individuelles entre les 14 topics de l'évaluation du point de vue de la f1-mesure. Elle illustre en particulier le fait que tous les topics ne présentent pas la même difficulté et que globalement, cette variation dans le degré de difficulté se fait ressentir à quelques exceptions près de la même façon pour les trois versions de REDUIT.

[9] Nous travaillons actuellement sur ce point sur la base des moyens utilisés pour la structuration thématique des profils.

10. Conclusion et perspectives

Nous avons présenté dans cet article une méthode pour la construction de résumés multi-documents permettant d'extraire d'un ensemble de documents ses parties les plus significatives relativement à un profil défini par un utilisateur. Cette méthode fait la distinction entre les parties de document en relation directe avec un profil et les parties de document représentant potentiellement des informations nouvelles liées à ce profil, ce qui présente un intérêt particulier pour des applications de veille technologique par exemple. Plus classiquement, les extraits similaires au sein de chacune de ces deux catégories sont regroupés et désignés par leur représentant le plus caractéristique.

L'implémentation et l'évaluation du système REDUIT ont permis d'illustrer de façon significative l'intérêt de l'approche que nous avons choisie. En particulier, les différentes évaluations menées ont démontré que la prise en compte explicite de l'hétérogénéité thématique des profils permet d'améliorer les résultats tant pour le filtrage de documents que l'extraction de passages. Ils montrent de plus qu'un enrichissement des profils à partir des documents filtrés s'avère également bénéfique pour cette même tâche d'extraction.

Pour prolonger ces résultats et les améliorer, nous nous focalisons sur plusieurs pistes. La première d'entre elles consiste à intégrer des outils de traitement linguistique des textes plus élaborés[10] afin d'extraire des termes complexes et des entités nommées et d'évaluer leur apport par rapport aux traitements actuels. Une autre piste importante est constituée par le problème de l'enrichissement des profils. Nous effectuons déjà un tel enrichissement par l'introduction des termes inférés mais comme nous l'avons vu, il reste à rattacher ces termes aux sous-thèmes. Par ailleurs, il est également intéressant d'examiner l'apport de dictionnaires de synonymes lors de la définition initiale des profils. Enfin, le processus de structuration des profils est capable de produire un enrichissement de nature thématique, déjà exploité dans le cadre de CLEF (Besançon *et al*., 2003) et que nous pourrions également utiliser pour REDUIT.

Toujours à propos des profils, il paraît intéressant d'offrir à l'utilisateur un autre mode de définition. Si celui-ci dispose d'un ensemble de documents relatifs aux thèmes du profil qu'il souhaite définir, l'utilisation conjointe, telle qu'elle a été décrite dans (Ferret, 1998), d'une segmentation thématique (cf. section 5.2) et d'une méthode de classification non supervisée permet de proposer à l'utilisateur une représentation des thèmes de ces documents qu'il pourra ensuite modifier manuellement s'il le souhaite.

Enfin, pour aller vers la construction d'un véritable résumé par extraction, il convient de passer de la sélection de segments de texte à celle d'unités textuelles plus fines, telles que les phrases, en allant au-delà de la simple heuristique que nous avons utilisée pour nos évaluations. Cette progression dans la granularité des unités

[10] Actuellement en cours de développement au sein du LIC2M.

extraites devrait par ailleurs nous permettre de nous inscrire plus facilement dans les cadres d'évaluation existants, en particulier les conférences DUC.

11. Bibliographie

Alonso, L., Fuentes M., « Collaborating discourse for Text Summarisation », *7th ESSLLI Student Session*, Trento, Italy, 2002.

Ando R. K., Boguraev B. K., Byrd R. J., Neff M. S., « Multi-summarization by Visualizing Topical Content », *ANLP/NAACL Workshop on Automatic Summarization*, 2000.

Barzilay R, Mc Keown, Kathleen R., Elhadad M,, « Information fusion in the context of multi-document summarization », *37th Annual Meeting of the ACL*, 1999.

Besançon R., de Chalendar G., Ferret O., Fluhr C., Mesnard O., Naets H., « The LIC2M's CLEF2003 System », *CLEF 2003 Workshop*, 2003.

Boros E, Kandor P., Neu D-J., « A clustering based approach to creating multi-document summaries », *ACM SIGIR'01 Workshop on Text Summarization*, 2001.

Châar S, Ferret O, Fluhr C., « Filtrage multi-document orienté par un profil utilisateur », *Conférence CIDE'5*, Hammamet, Octobre 2002.

Châar S., « Extraction de segments thématiques pour la construction de résumé multi-document orienté par un profil utilisateur », *Récital 2003*, Batz-sur-Mer, France, 2003.

Châar S, Ferret O, Fluhr C., « Génération de résumé multi-document guidé par un profil utilisateur », *Conférence SETIT'2004*, Sousse, Tunisie, 2004.

de Chalendar G., Dalmas T., Elkateb-Gara F., Ferret O., Grau B., Hurault-Plantet M., Illouz G., Monceaux L., Robba I,. Vilnat A., « The question answering system QALC at LIMSI: experiments in using Web and WordNet », *11th Text Retrieval Conference (TREC-2002)*, 2003.

Choi F., « Advances in domain independent linear text segmentation », *NAACL'00,* p. 26-33, 2000.

Ferret O., « Filtrage thématique d'un réseau de collocations », *TALN 2003*, p. 347-352, 2003.

Ferret O., « Using collocations for topic segmentation and link detection », *COLING 2002*, p. 260-266, 2002.

Ferret O., Grau B., « A Thematic Segmentation Procedure for Extracting Semantic Domains from Texts », *ECAI'98*, p. 155-159, 1998.

Fluhr C., « SPIRIT : un système d'exploration de données textuelles », *Le Traitement Informatique des Corpus Textuels*, INALF, 1994.

Fukusima T, Okumura M., « Text Summarization Challenge: Text Summarization Evaluation at NTCIR Workshop2. » *In Proceedings of the Second NTCIR Workshop on Research in Chinese & Japanese Text Retrieval and Text Summarization, NII,* Tokyo, Japan, 2001.

Goldstein J., Mittal V., Kantrowitz M., Carbonell J., « Multi-Document Summarization By Sentence Extraction », *ANLP/NAACL Workshop on Automatic Summarization*, 2000.

Hearst M., « TextTiling: Segmenting Text into Multi-paragraph Subtopic Passages », *Computational Linguistics*, Vol. 23, n° 1, p. 33-64, 1997.

Hull D., Robertson S., « The TREC-8 filtering Track Final Report, *8th Text Retrieval Conference (TREC-8)* », p. 35-55, 2000.

Kan M-Y., Klavans J., McKeown K. R., « Linear segmentation and segment significance », *6th Workshop on Very Large Corpora*, p. 197-205, 1998.

Kraaij W., Spitters M., van der Heijden M., « Combining a mixture language model and Naive Bayes for multi-document summarisation », *ACM SIGIR'01 Workshop on Text Summarization*, 2001.

Jacquemin C., Variation terminologique : Reconnaissance et acquisition automatiques de termes et de leurs variantes en corpus, Mémoire d'habilitation à diriger des recherches en informatique fondamentale, Université de Nantes, 1997.

Lehmam A., «Le résumé automatique à Fragments indicateurs : RAFI », Thèse de doctorat de l'Université de Nancy II, 1995.

Lin C-Y., Hovy E.H., « Automatic Evaluation of Summaries Using N-gram Co-occurrence Statistics », *HLT-NAACL 2003*, Edmonton, Canada, 2003.

Lin C-Y., Hovy E.H., « Manual and Automatic Evaluations of Summaries », *Document Understanding Conference (DUC-02) Workshop on Multi-Document Summarization Evaluationtof the ACL-02 Conference*, Philadelphia, U.S.A, 2002.

Lin X., « Map displays for information retrieval », *Information Processing & Management*, vol. 29, n° 1, 1993.

Mani, I., Bloedorn, E., « Summarizing similarities and differences among documents », *Information Retrieval*, vol. 1, n° 1, p. 1-23, 1999.

Mani I, House D, Klein G, Hirschman L, Obrst L., « The TIPSTER SUMMAC summarization Text evaluation, Final Report », *MITRE technical report*, October 1998.

McKeown K. R., Barzilay R., Evans D., Hatzivassiloglou V., Kan M-Y., Schiffman B., Teufel S., « Columbia Multi-document Summarization: Approach and Evalution », *ACM SIGIR'01 Workshop on Text Summarization*, 2001.

Minel J-L., Desclés J-P.; « Résumé Automatique et Filtrage des textes », *Ingénierie des langues*, Paris, Editions Hermès, 2000.

Minel J-L., *Filtrage sémantique Du résumé automatique à la fouille de texte*, Paris, Editions Hermès Lavoisier, 2002.

Olsen K., Korfhage R, Sochats K., Spring M., Williams J., « Visualization of a document collection: The VIBE System », *Information Processing & Management*, vol. 29, n° 1, 1993.

Over P., « Introduction to DUC-2001: an Intrinsic Evaluation of Generic News Text Summarization Systems », *ACM SIGIR'01 Workshop on Text Summarization*, 2001.

Radev, D. R., Hovy E.H., McKeown, K. R., « Introduction to the Special Issue on Summarization », *Computational Linguistics*, vol. 28, n° 4, p. 399-408, 2002.

Radev, D. R., Hongyan, J., Malgorzata, B., « Centroid-based summarization of multiple documents: Sentence extraction, utility-based evaluation, and user studies », *ANLP/NAACL Workshop on Summarization*, Seattle, U.S.A, 2000.

Radev, D. R., McKeown, K. R., « Generating natural language summaries from multiple on-line sources », *Computational Linguistics*, vol. 24, n°3, p. 469-500, 1998.

Saggion, H., Lapalme, G., « Generating Indicative-Informative Summaries with SumUM », *Computational Linguistics*, vol. 28, n° 4, p. 497-526, 2002.

Sparck Jones, K., Galliers, G., *Evaluating Natural Language Processing Systems: An Analysis and Review*, Lecture Notes in Artificial Intelligence n°1083, Springer, 1995.

Stein G. C., Bagga A., Wise B. G., «Multi-Document Summarization: Methodologies and Evaluations», *TALN'00*, p. 337-346, 2000.

Utiyama M., Isaharan H., « A Statistical Model for Domain-Independent Text Segmentation », *ACL 2001*, p. 491-498, 2001.

Voorhees E., Tice D., « The TREC-8 Question Answering Track », *Language Resources and Evaluation Conference*, 2000.

White M., Cardie C., Ng V., Wagstaff K., McCullough D., « Detecting Disrepancies and Improving Intelligibility: Two Preliminary Evaluations of RIPTIDES », *ACM SIGIR'01 Workshop on Text Summarization*, 2001.