Approches endogène et exogène pour améliorer la segmentation thématique de documents

Olivier Ferret

CEA-LIST/LIC2M
18, route du Panorama – B.P. 6, F-92265 Fontenay-aux-Roses Cedex
ferreto@zoe.cea.fr

RÉSUMÉ. *La segmentation thématique de documents a fait l'objet d'un nombre important de travaux dont il n'est pas toujours facile de dégager des conclusions claires, en particulier en ce qui concerne l'utilisation de connaissances. Dans cet article, nous proposons d'examiner deux voies se situant dans le même cadre pour améliorer une méthode de segmentation fondée sur la simple récurrence lexicale. La première est endogène. Elle exploite la similarité distributionnelle des mots au sein des documents pour en découvrir les thèmes. Ces thèmes sont ensuite utilisés pour faciliter l'identification des similarités thématiques entre unités de discours. La seconde réalise le même but en faisant appel à une ressource externe, en l'occurrence un réseau de cooccurrences lexicales construit à partir d'un large corpus. Ces deux approches sont également combinées. Une évaluation de ces deux approches et de leur combinaison est réalisée dans un même cadre et illustre l'intérêt de cette combinaison.*

ABSTRACT. *Topic segmentation was addressed by a large amount of work from which it is not easy to draw conclusions, especially about the need for knowledge. In this article, we propose in the same framework two methods for improving the results of a topic segmenter based on lexical reiteration. The first one is endogenous and exploits the distributional similarity of the words of a document for discovering its topics. These topics are then used to facilitate the detection of topical similarity between discourse units. The second approach achieves the same goal by relying on an external resource, that is a network of lexical co-occurrences built from a large corpus. These two approaches are also combined. An evaluation of these approaches and their combination is performed in a reference framework and shows the interest of this combination.*

MOTS-CLÉS : *analyse thématique, segmentation thématique, découverte de thèmes, cooccurrences lexicales.*

KEYWORDS: *topic analysis, topic segmentation, topic discovery, lexical co-occurrences.*

TAL. Volume 47 – n° 2/2006, pages 111 à 135

1. Introduction

Le problème auquel nous nous intéressons dans cet article est celui de la segmentation thématique, problème consistant à découper linéairement un document en une suite de segments thématiquement homogènes. Cette partie de l'analyse du discours a fait l'objet de nombreux travaux ces dernières années à la suite notamment de celui de Hearst (Hearst, 1994). Sur le plan applicatif, elle intervient dans différentes tâches d'accès à l'information thématique dont la plus notable est le résumé automatique, comme l'illustrent des travaux tels que (Barzilay et Elhadad, 1997), (Boguraev et Neff, 2000) ou (Châar *et al.*, 2004). L'intérêt porté à la segmentation thématique s'est manifesté en particulier par sa présence au travers d'une tâche dédiée dans le cadre des évaluations Topic Detection and Tracking (TDT), consacrées plus généralement à différentes dimensions de l'analyse thématique à la fois de documents écrits et de transcriptions de parole.

Sur un plan plus méthodologique, les travaux sur la détection des changements de thème dans un texte ont emprunté deux voies :

– l'identification de marques, en particulier linguistiques, caractéristiques des changements de thème ;

– la détection des changements du contenu du discours.

Il faut en préambule remarquer que la première approche a été globalement moins étudiée que la seconde. Une explication possible de cette situation est que s'attacher au contenu du discours est en apparence[1] toujours réalisable alors que la seconde approche est limitée aux discours présentant des changements de thème caractérisés par des marques spécifiques. Or ces marques ne sont en pratique pas très fréquentes et sont par ailleurs assez dépendantes du type de discours considéré. Lorsqu'elles sont présentes, elles ont néanmoins l'avantage de localiser les changements de thème avec précision ce qui n'est, à l'inverse, pas toujours le cas des méthodes fondées sur le contenu du discours.

Les deux approches sont donc plus complémentaires que concurrentes. C'est d'ailleurs ce que l'on observe dans les travaux existants puisqu'une part importante de ceux exploitant des marques linguistiques de segmentation le font en conjonction avec des indices relatifs au contenu du discours. Ainsi dans (Passonneau et Litman, 1997), un de ces premiers travaux, des marqueurs discursifs sont associés à la détection des ruptures dans les chaînes coréférentielles. Plus récemment, (Tür *et al.*, 2001), (Galley *et al.*, 2003) ou (Couto *et al.*, 2004) ont conjugué l'utilisation de marqueurs discursifs et de méthodes de détection des changements du contenu du discours s'appuyant sur le lexique. Une autre caractéristique notable de ces travaux est que bon nombre d'entre eux ont été réalisés dans le contexte du discours oral. De ce fait, une partie des marques exploitées pour la segmentation thématique sont spécifiques de la parole :

1. En apparence seulement car, ainsi que nous l'avons illustré dans (Ferret *et al.*, 1998), l'efficacité des méthodes fondées sur le contenu peut être très variable selon le type de texte considéré et les connaissances qu'elles utilisent.

chevauchement des locuteurs, temps de pause ou marques prosodiques. Ce type de discours offre en effet une plus grande richesse de marques exploitables que le discours écrit et, à l'inverse, il ne se prête pas nécessairement bien à la détection des changements de contenu car les segments thématiques y sont souvent de petite taille.

Dans le cadre de cet article, nous nous focaliserons plus spécifiquement sur des documents textuels et sur les moyens d'améliorer une segmentation dirigée par le contenu. Mais cette étude est à replacer dans le contexte plus large que nous avons présenté dans (Couto *et al.*, 2004) où ce type de segmentation coopère avec une segmentation exploitant les marques linguistiques associées à la notion de cadre du discours (Charolles, 1997).

2. Une segmentation dirigée par le contenu

Un des critères permettant d'appréhender les travaux dans le domaine de la segmentation thématique dirigée par le contenu est le type de connaissances sur lesquelles ils reposent. La plupart d'entre eux s'appuient sur les seules caractéristiques intrinsèques des documents : la récurrence lexicale dans le cas de (Hearst, 1994), (Choi, 2000), (Utiyama et Isahara, 2001) ou (Galley *et al.*, 2003) ; la réitération des multi-termes et des entités nommées dans le cas de (Kan *et al.*, 1998). Ces méthodes ne faisant pas appel à des connaissances externes, elles ne sont pas limitées à un champ thématique particulier. En revanche, leur application est restreinte à un certain type de documents : la récurrence lexicale n'est ainsi un indice thématique fiable que si les concepts du document considéré ne sont pas exprimés sous des formes trop diverses (synonymes, hyperonymes, etc.).

Une des pistes suivies par un certain nombre de systèmes pour surmonter ces limitations est d'exploiter des connaissances sur les relations de cohésion lexicale, connaissances qui présentent elles aussi l'avantage de ne pas dépendre d'un domaine particulier. Elles prennent la forme d'un réseau lexical construit à partir d'un dictionnaire dans (Kozima, 1993), d'un thésaurus dans (Morris et Hirst, 1991) ou encore d'un large ensemble de cooccurrences lexicales dans (Ferret, 1998), (Kaufmann, 1999) ou (Choi *et al.*, 2001). D'une certaine façon, ces connaissances permettent aux systèmes de segmentation thématique de détecter les récurrences à un niveau plus conceptuel. Cependant, leur nature lexicale et leur absence de structuration thématique explicite font que les systèmes qui les utilisent sont parfois mis en échec par l'ambiguïté sémantique des mots ou par l'impossibilité d'identifier des relations comme spécifiquement thématiques.

La solution la plus simple concernant ce dernier point est la possibilité d'exploiter des connaissances sur les thèmes susceptibles d'être rencontrés dans les documents analysés. C'est en particulier l'approche retenue par les participants aux évaluations TDT qui construisent une représentation des thèmes rencontrés à partir des documents exemples fournis. Cette approche est typiquement représentée par le travail décrit dans (Yamron *et al.*, 1998) et se retrouve dans une partie de (Beeferman *et al.*, 1999) ou de

(Tür *et al.*, 2001). Le travail de Bigi (Bigi *et al.*, 1998) s'inscrit dans la même perspective mais en se focalisant sur des thèmes plus larges que ceux considérés dans TDT tandis que (Ferret et Grau, 2000) opère avec des représentations des thèmes construites de façon non supervisée. Plus généralement, les connaissances thématiques apprises par ces systèmes leur permettent une plus grande précision mais restreignent parallèlement leur champ d'action à des documents portant sur une certaine thématique.

Enfin, des systèmes hybrides combinant différentes approches parmi celles présentées ci-dessus ont également été développés et ont prouvé leur intérêt : (Jobbins et Evett, 1998) associe ainsi la récurrence lexicale, l'utilisation de cooccurrences et celle d'un thésaurus ; (Beeferman *et al.*, 1999) et (Tür *et al.*, 2001) s'appuient à la fois sur une modélisation statistique des thèmes et sur l'utilisation de marques discursives ; (Galley *et al.*, 2003) exploite conjointement la récurrence lexicale et des marques discursives.

Le travail que nous présentons dans cet article se situe dans une perspective proche de celle de (Jobbins et Evett, 1998). Les changements de thème au sein d'un document y sont repérés en détectant les changements de son contenu au niveau lexical et, dans ce cadre, nous nous intéressons particulièrement à la comparaison entre plusieurs approches pour réaliser une telle détection : de façon purement endogène, c'est-à-dire en n'utilisant que le document traité, ou en faisant appel à des sources de connaissances externes au document, approche que nous qualifierons d'exogène.

3. Problématique

Les algorithmes de segmentation s'inscrivant dans la filiation plus ou moins directe de l'algorithme *TextTiling* de Hearst prennent comme point de départ une représentation des documents sous la forme d'une séquence d'unités de discours. Dans le cas de documents écrits, il s'agit généralement de phrases, approche que nous avons également adoptée dans notre travail. Chaque unité est transformée en un vecteur de mots suivant les principes du modèle *Vector Space* (Salton et McGill, 1983). La similarité entre deux unités peut ainsi être évaluée en faisant appel à une mesure de similarité vectorielle, comme la mesure *cosinus* par exemple. Dans ce contexte, une telle mesure est considérée comme représentative de la proximité thématique des unités impliquées. Compte tenu des caractéristiques du modèle *Vector Space*, ce principe s'étend à des regroupements d'unités, comme des segments de document. La détection des changements de thème s'identifie alors à la détection des zones dans lesquelles la similarité entre unités ou entre regroupements d'unités est faible.

Cette vue d'ensemble souligne le rôle central de l'évaluation de la similarité entre les unités de discours dans ce type de méthode. Lorsque aucune connaissance externe n'est utilisée, cette similarité ne repose que sur la réitération lexicale. Mais il est possible d'y intégrer la prise en compte de relations sémantiques entre les mots. Cette prise en compte peut être faite par l'utilisation de relations sémantiques clairement identifiées : (Jobbins et Evett, 1998) puise ces relations dans le *Roget Thesaurus*. Cette

ressource est également utilisée dans (Morris et Hirst, 1991), où la similarité entre les segments de discours est évaluée de manière plus indirecte par le biais des chaînes lexicales qu'ils partagent. La même approche se retrouve dans (Stokes *et al.*, 2002) mais avec le réseau lexico-sémantique *WordNet* (Fellbaum, 1998) comme ressource sémantique de référence.

Ces relations peuvent également avoir un statut plus implicite, comme c'est le cas avec les cooccurrences lexicales. De telles cooccurrences, qui recouvrent des relations plus souvent de nature syntagmatique que paradigmatique, sont ainsi utilisées dans le cadre de la segmentation thématique par (Jobbins et Evett, 1998), (Ferret, 2002) ou encore par (Dias et Alves, 2005) pour améliorer la détection des similarités thématiques entre unités de discours. La prise en compte des relations sémantiques entre ces unités peut être effectuée de façon encore plus indirecte lorsqu'elle est menée à bien par le biais d'une projection dans un espace sémantique. C'est ce qui est réalisé dans le système CWM (Choi *et al.*, 2001), une variante de l'algorithme C99 dans laquelle chaque mot est remplacé par sa représentation dans un espace issu de l'*Analyse sémantique latente* (Landauer *et al.*, 1998). Le même type de démarche se retrouve dans (Ponte et Croft, 1997) et (Caillet *et al.*, 2004) où un mot est représenté par sa proximité par rapport à un ensemble de concepts de référence construits automatiquement à partir d'un corpus, en utilisant dans le premier cas la méthode *Local Context Analysis* (Xu et Croft, 1996) et dans le second cas, l'algorithme de classification *X-means* (Pelleg et Moore, 2000).

Dans cet article, nous proposons de prendre comme point de départ une méthode de segmentation thématique fondée sur la récurrence lexicale et d'examiner comment les résultats de cette méthode peuvent être augmentés en améliorant la détection des similarités thématiques entre unités de discours. Cette amélioration est réalisée selon une approche endogène et selon une approche exogène, puis en combinant ces deux voies. En utilisant un cadre de travail unique pour tester ces différentes modalités, notre objectif est ainsi d'obtenir une vision plus claire de l'intérêt de ces différentes approches, vision qu'il n'est pas toujours facile de dégager des travaux existants.

4. Une segmentation thématique fondée sur la récurrence lexicale: F06

La méthode de segmentation thématique proposée par Hearst dans (Hearst, 1994), *TextTiling*, se décompose en trois grandes parties :

– le prétraitement linguistique des documents ;

– l'évaluation de la cohésion lexicale au sein du document ;

– l'identification des changements de thème.

La méthode que nous proposons ici, F06, reprend ces trois grandes étapes mais avec des modalités un peu différentes de celles adoptées par Hearst. Le prétraitement linguistique découpe les documents en phrases et représente chacune d'elles comme la séquence de ses mots pleins normalisés, c'est-à-dire ses noms (noms com-

muns et noms propres), ses verbes et ses adjectifs. Il est réalisé par l'outil *TreeTagger* (Schmid, 1994), qui conjugue pour ce faire tokenization, analyse morphologique et étiquetage morpho-syntaxique. La normalisation des mots est la première façon, réalisée au niveau morphologique, de favoriser la détection de la similarité entre segments de document. L'évaluation de la cohésion lexicale s'appuie comme dans *TextTiling* sur l'utilisation d'une fenêtre glissante de taille fixe. Cette fenêtre se déplace sur le texte de phrase en phrase. À chaque station de cette fenêtre, la cohésion lexicale est évaluée en son sein et affectée à la fin de phrase sur laquelle elle est centrée. Cette évaluation est réalisée suivant le principe proposé dans (Jobbins et Evett, 1998) : la cohésion est mesurée par l'application du coefficient de *Dice* entre les vecteurs représentant les deux moitiés de la fenêtre glissante. Plus précisément, si F_g désigne le vocabulaire de la moitié gauche de la fenêtre et F_d, celui de la moitié droite de cette même fenêtre, la cohésion au sein de celle-ci est donnée par :

$$cohésion\_réc(x) = \frac{2 \cdot card(F_g \cap F_d)}{card(F_g) + card(F_d)} \quad [1]$$

Cette mesure a été utilisée plutôt que la mesure *cosinus* retenue pour *TextTiling* car sa définition ensembliste rend son extension plus facile pour intégrer d'autres relations que la simple récurrence lexicale, à l'instar de (Jobbins et Evett, 1998). La cohésion est ainsi évaluée pour chaque frontière inter-phrastique du document considéré et le résultat global est une courbe de cohésion couvrant l'ensemble du document, comme l'illustre la figure 1.

La troisième partie de l'algorithme s'inspire, quant à elle, de son homologue dans le système LCseg (Galley *et al.*, 2003). Elle comprend elle-même trois étapes :

– le calcul d'un score évaluant la probabilité pour chaque minimum de la courbe de cohésion de correspondre à un changement de thème ;

– la suppression des candidats segments de trop petite taille ;

– la sélection des bornes de segments thématiques.

Le calcul du score initial d'un minimum commence par la recherche de la paire de maxima g et d qui l'entourent. En notant, $CL(x)$ la valeur de la cohésion lexicale à la position x, le score d'un minimum m est donné par :

$$score(m) = \frac{CL(g) + CL(d) - 2 \cdot CL(m)}{2} \quad [2]$$

Ce score, compris entre 0 et 1, est d'autant plus élevé que la différence entre le minimum considéré et les maxima qui l'entourent est plus importante. Il favorise ainsi comme changements de thème potentiels les minima caractérisés par une chute très nette de la cohésion lexicale. La suppression des candidats segments trop petits s'effectue, quant à elle, par une simple comparaison par rapport à un seuil de référence : les minima se trouvant à P phrases au plus du minimum qui les précèdent (P étant égal à 2 dans le cas des expérimentations de la section 7) sont éliminés en tant que possibles changements de thèmes. Finalement, la sélection des bornes de segments

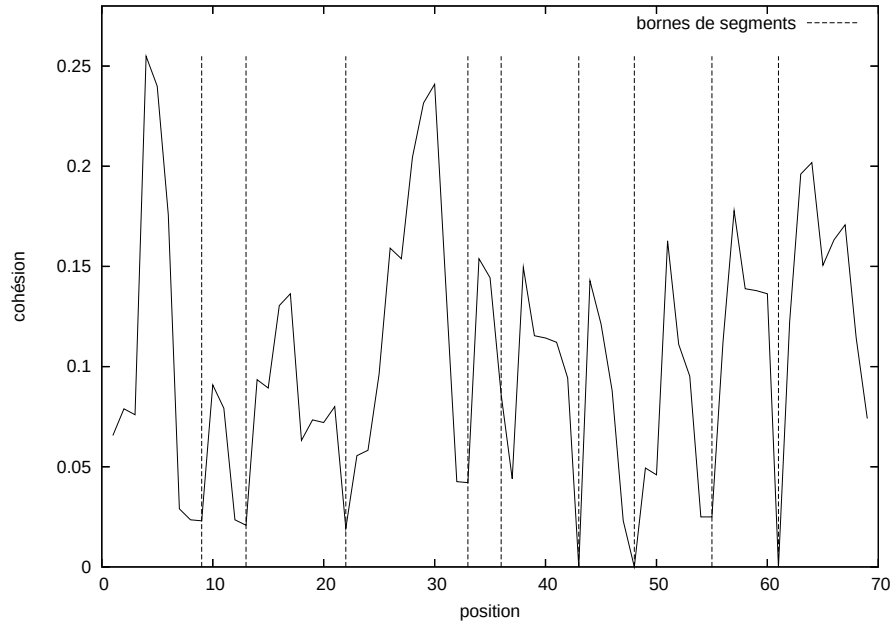

Figure 1. *Courbe de cohésion de F06 pour un document du corpus de la section 7.1*

thématiques est réalisée par l'utilisation d'un seuil s'adaptant à la distribution des scores des minima. Un minimum m est ainsi retenu comme borne de segment si :

$$score(m) > \mu - \alpha \cdot \sigma \quad [3]$$

où μ correspond à la moyenne des scores de minima, σ, à l'écart-type de ces scores et α, à un coefficient de modulation.

5. Une approche endogène pour améliorer la segmentation thématique

5.1. *Principes*

Pour améliorer la méthode F06 tout en s'inscrivant dans le cadre qu'elle définit, nous avons d'abord opté pour une solution ne faisant pas appel à des ressources externes au document traité (Ferret, 2006). Plus précisément, l'idée est d'exploiter les relations de proximité entre mots pouvant être extraites à l'échelle du document pour aider à identifier les proximités thématiques au niveau plus local des unités de dis-

cours[2]. Dans un premier temps, nous identifions ainsi les thèmes de chaque document à segmenter en procédant à une classification non supervisée de son vocabulaire. Cette classification s'appuie sur les cooccurrences enregistrées au sein du document entre les mots de ce vocabulaire. À la suite de cette phase, chaque thème est représenté par un sous-ensemble du vocabulaire du document. Lors de la segmentation, l'évaluation de la similarité entre deux segments repose d'abord sur la proportion des mots communs à ces deux segments, mais elle prend aussi en compte la proportion de leurs mots appartenant à un même thème parmi ceux identifiés précédemment dans le document. Ainsi, deux segments peuvent être jugés proches parce qu'évoquant un même thème sans pour autant partager un nombre important de mots. Plus globalement, cette prise en compte des thèmes des documents vise à faire diminuer le taux de détection de « faux » changements de thème.

5.2. *Découvrir les thèmes d'un document*

Pour découvrir les thèmes d'un document sans utiliser de connaissances *a priori*, nous faisons l'hypothèse que les mots les plus représentatifs de chaque thème apparaissent dans des contextes similaires. Étant donné cette hypothèse, nous collectons les cooccurrents de chaque mot du document possédant une fréquence d'occurrence minimale, nous évaluons la similarité deux à deux de ces mots en nous appuyant sur leurs cooccurrents pour finalement construire une représentation des thèmes du document en appliquant à ces mots une méthode de classification non supervisée.

5.2.1. *Évaluer la proximité thématique des mots d'un texte*

La découverte des thèmes d'un document prend comme point de départ le document à l'issue de l'application du prétraitement linguistique décrit à la section 4. Les mots de plus faible fréquence sont ensuite filtrés et les cooccurrents des mots restant du document sont collectés suivant les principes décrits dans (Church et Hanks, 1990), en enregistrant les cooccurrences au sein d'une fenêtre de taille fixe déplacée sur tout le document prétraité. À la suite de cette étape, chaque mot sélectionné est représenté par le vecteur de ses fréquences de cooccurrence avec les autres mots du document. La similarité deux à deux de tous les mots sélectionnés est alors évaluée pour constituer leur matrice de similarité. Classiquement, nous appliquons la mesure *cosinus* entre les vecteurs représentant ces mots pour réaliser cette évaluation.

5.2.2. *De la proximité des mots aux thèmes d'un document*

L'étape finale de la découverte des thèmes d'un document est la classification non supervisée de ses mots à partir de la matrice de similarité construite précédemment. Nous nous appuyons pour cette tâche sur une adaptation de l'algorithme *Shared Nea-*

2. Une telle approche n'est bien sûr applicable que pour des documents d'une taille suffi sante. Néanmoins, l'évaluation de la section 7.3 montre que cette taille n'est pas nécessairement importante puisque les documents de cette évaluation comportent entre 65 et 70 phrases.

rest Neighbors (SNN), décrit dans (Ertöz *et al.*, 2001). Cet algorithme s'accorde particulièrement avec nos besoins dans la mesure où il détermine automatiquement le nombre de classes, dans notre cas le nombre de thèmes d'un document, et qu'il détecte les éléments ne s'accordant pas avec les classes qu'il constitue. Ce dernier point est particulièrement important étant donné que tous les mots pleins d'un document ne sont pas spécifiques de ses thèmes.

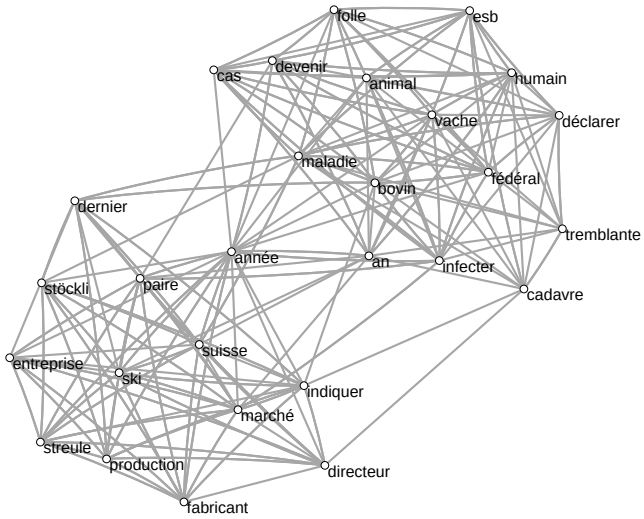

Figure 2. *Graphe de similarité des mots d'un document après l'étape 1*

L'algorithme SNN s'inscrit dans la mouvance des algorithmes ramenant le problème de la classification à celui de la détection de composantes de forte densité dans un graphe de similarité. Dans un tel graphe, chaque nœud représente un élément à classer et une arête relie deux nœuds lorsque la similarité entre les éléments qu'ils représentent est non nulle. Dans notre cas, le graphe de similarité est construit directement à partir de la matrice de similarité des mots du document. Dans son principe général, l'algorithme SNN comporte deux grandes étapes : la première vise à mettre en évidence les éléments les plus représentatifs de leur voisinage en masquant les relations les moins importantes du graphe de similarité. Ces éléments constituent les embryons des futures classes, formées dans un second temps en agrégeant les autres éléments à ceux sélectionnés lors de la première phase. L'algorithme SNN tel que nous l'avons adapté à la découverte des thèmes d'un document se décompose plus précisément comme suit :

1) Éclaircissement du graphe de similarité. Pour chaque mot sélectionné du document, seules les arêtes en direction des k (k=10 en l'occurrence) plus proches mots sont conservées. La figure 2 illustre la forme du graphe résultant pour un document de

notre corpus d'évaluation (cf. section 7.1) faisant référence à deux thèmes (la maladie de la vache folle d'une part, la fabrication de ski d'autre part).

2) Construction du graphe des plus proches voisins partagés. Cette étape consiste à remplacer dans le graphe éclairci la valeur portée par chaque arête par le nombre de voisins directs que les deux mots reliés par l'arête ont en commun.

3) Calcul de la distribution en liens forts des cooccurrents. L'objectif de cette étape est, comme lors de l'étape 1, de procéder à une sorte d'éclaircissement. Il s'agit de repérer les mots du document autour desquels s'organisent un ensemble d'autres mots, *i.e.* des germes de thème, mais aussi de repérer ceux qui sont visiblement sans connexion véritable avec les autres. Pour ce faire, un seuil minimum est fixé concernant le nombre de voisins partagés par deux mots du document, seuil au-dessus duquel on considère les deux mots comme fortement liés. On caractérise ensuite chaque mot du document par le nombre de liens forts qu'il possède.

4) Détermination des germes de thème et élimination du bruit. Les germes de thème et les mots du document laissés de côté sont déterminés par simple comparaison de leur nombre de liens forts par rapport à un seuil.

5) Construction des thèmes. Cette étape consiste principalement à associer aux germes de thème trouvés à l'étape précédente les mots du document non déjà sélectionnés comme germe de thème ou bruit pour former des classes représentant les thèmes du document. Pour associer un mot à un germe de thème, la force du lien qui les unit doit être supérieure à un seuil. Si plusieurs germes de rattachement sont possibles, le choix se porte sur celui avec lequel la force du lien est la plus grande. Par ailleurs, cette étape est aussi l'occasion de rassembler plusieurs germes de thème considérés comme trop proches pour former des thèmes distincts : le rattachement des mots fait donc également intervenir les germes de thème.

6) Élimination des thèmes non représentatifs. Lorsque le nombre de mots rassemblés par un thème est trop petit, ce thème est considéré comme non représentatif et il est supprimé. Ses mots sont alors rattachés à l'ensemble des mots non affectés à l'issue de l'étape précédente.

7) Élargissement des thèmes. À l'issue des étapes précédentes, un nombre plus ou moins important de mots du document n'ayant pas été considérés comme du bruit se retrouvent néanmoins sans affectation à un thème. Ce nombre dépend bien entendu de la sévérité du seuil de rattachement à un germe de thème mais l'objectif étant de former des classes homogènes, celle-ci doit être nécessairement assez forte. Néanmoins, il est également intéressant que les thèmes puissent être décrits de la façon la plus complète et la plus précise possible. Les thèmes à ce stade étant caractérisés de façon plus sûre qu'à l'issue de l'étape 4, il est possible de leur rattacher des mots du document dont la force de lien avec leurs constituants est plus faible de façon plus sûre.

Par rapport à l'algorithme SNN décrit dans (Ertöz *et al.*, 2001), nous avons ajouté l'étape 6 car nous avons observé que malgré la possibilité de fusionner des classes lors de l'étape 5, certains thèmes restent divisés en plusieurs classes. Dans un nombre significatif de cas, le thème « divisé » se répartit entre une ou plusieurs classes ne

regroupant que 3 à 4 mots et une classe de plus large ampleur. Plus précisément, les germes de thème de ces classes « minoritaires » n'ont pas pu être rattachés à la classe « majoritaire » alors que la plupart des mots qui leur étaient liés s'y sont rattachés. Pour regrouper ces classes « minoritaires » avec la classe la plus importante, nous avons choisi de laisser l'algorithme SNN le faire en détruisant ces classes et en remettant leurs éléments dans l'ensemble des cooccurrents non rattachés. La dernière étape de l'algorithme permet alors de rattacher ces cooccurrents à la classe « majoritaire » dans la plupart des cas. De plus, ce mécanisme permet d'obtenir une plus grande stabilité des thèmes formés lorsque les paramètres de l'algorithme sont modifiés.

Concernant les différents seuils de l'algorithme, nous avons opté pour un mode unique de fixation s'adaptant à la distribution des valeurs considérées : chaque seuil est exprimé comme un quantile de ces valeurs. Pour le seuil de détermination des germes de thème et de celui de définition du bruit (cf. étape 4), il s'agit d'un quantile du nombre de liens forts des mots du document. Pour le seuil définissant la notion de lien fort (cf. étape 3), celui de rattachement des mots aux germes (cf. étape 5) et celui de rattachement des mots aux thèmes (cf. étape 7), le quantile est appliqué à la force des liens entre les mots dans le graphe des plus proches voisins partagés.

À l'issue de l'application de l'algorithme SNN, un ensemble de thèmes, éventuellement vide si le niveau global de récurrence lexicale est trop faible, est donc associé au document à segmenter, chacun d'entre eux étant défini par un sous-ensemble du vocabulaire de ce document.

5.3. *Utiliser les thèmes découverts pour la segmentation thématique*

Le cœur de l'algorithme présenté à la section 4 est l'évaluation de la cohésion à l'intérieur de la fenêtre glissante, donnée par l'équation 1. Cette évaluation est également son point faible car le terme ne s'appuie que sur la notion de récurrence lexicale. Ainsi, deux mots différents appartenant respectivement à F_g et F_d, mais faisant partie du même thème, ne peuvent en aucune façon contribuer à identifier une éventuelle similarité thématique entre les deux volets de la fenêtre.

L'algorithme F06T reprend les principes de F06 mais en étendant l'évaluation de la cohésion au sein de la fenêtre glissante pour y ajouter la prise en compte des proximités thématiques entre les mots. Les thèmes de référence sont bien entendu les thèmes du document découverts par la méthode exposée à la section 5.2. Dans cette version étendue, l'évaluation de la cohésion au sein de la fenêtre glissante s'articule en trois étapes :

– le calcul de la cohésion s'appuyant sur la seule récurrence lexicale ;

– la détermination du ou des thèmes de la fenêtre ;

– le calcul de la cohésion s'appuyant sur les thèmes de la fenêtre et sa combinaison avec la cohésion fondée sur la récurrence lexicale.

La première étape est strictement identique au calcul de cohésion réalisé dans F06. La deuxième a pour objectif de restreindre les thèmes utilisés lors de la dernière étape aux thèmes véritablement représentatifs du contenu de la fenêtre glissante, c'est-à-dire représentatifs du contexte courant du discours. Ce problème de représentativité est particulièrement sensible dans les zones de changement de thème. Pour détecter l'instabilité thématique qui les caractérisent, il ne faut pas, en effet, amplifier les manifestations des thèmes environnants. Pour ce faire, un thème est considéré comme représentatif du contenu de la fenêtre seulement s'il s'apparie avec chacun des deux volets de cette fenêtre. En pratique, cet appariement est évalué en appliquant la mesure *cosinus* entre le vecteur représentant une moitié de la fenêtre et le vecteur représentant le thème[3]. Le thème est jugé représentatif de la fenêtre lorsque la valeur de cette mesure est supérieure à un seuil fixé a priori (égal à 0,1 dans les expérimentations de la section 7). Par ailleurs, nous ne cherchons pas à n'associer à la fenêtre qu'un seul thème du document. Comme la découverte des thèmes se fait de façon non supervisée et sans connaissances externes, il se peut qu'une thématique du document se retrouve répartie sur plusieurs représentations de thème. L'évaluation de la similarité avec la fenêtre est donc réalisée avec tous les thèmes du document à chaque nouvelle station de la fenêtre et tous ceux remplissant la condition mentionnée ci-dessus sont retenus pour l'étape suivante.

Cette dernière étape consiste en premier lieu à déterminer pour chaque volet de la fenêtre glissante le nombre de ses mots présents dans l'un des thèmes représentatifs du contenu de cette fenêtre. La cohésion de la fenêtre évaluée sur la base des thèmes du document est ensuite obtenue grâce à l'équation [4], qui rapporte le niveau de représentation des thèmes de la fenêtre dans ses deux parties au nombre total de mots de la fenêtre :

$$cohésion\_thm(x) = \frac{card((F_g \cap T_f) - M_{réc}) + card((F_d \cap T_f) - M_{réc})}{card(F_g) + card(F_d)} \quad [4]$$

où T_f désigne l'union des représentations des thèmes associés à la fenêtre glissante et $M_{réc} = F_g \cap F_d$, c'est-à-dire l'ensemble des mots de la fenêtre sur lesquels s'appuie l'évaluation de la cohésion fondée sur la récurrence lexicale.

En final, la cohésion globale au sein de cette fenêtre est donnée par l'addition de $cohésion\_réc(x)$, la cohésion évaluée sur la base de la récurrence lexicale (cf. équation [1]), et de $cohésion\_thm(x)$, la cohésion évaluée à partir des thèmes du document (cf. équation [4]). On notera la présence du terme $M_{réc}$ dans $cohésion\_thm(x)$ pour éviter d'accorder au niveau de la cohésion globale une importance excessive aux mots récurrents faisant partie des thèmes associés à la fenêtre.

3. Dans le vecteur représentant un thème, tous les mots du thème se voient attribuer un poids égal à 1. Dans celui représentant une moitié de la fenêtre, ce poids est égal au nombre d'occurrences du mot dans la fenêtre.

6. Une approche exogène pour améliorer la segmentation thématique

La seconde approche que nous avons testée afin d'améliorer les performances de F06 est de faire appel à des connaissances extérieures au document traité. À l'instar de travaux évoqués à la section 3, comme (Jobbins et Evett, 1998), (Ferret, 2002) ou (Dias et Alves, 2005), nous avons choisi de nous appuyer sur un réseau de cooccurrences lexicales, source de connaissances rendant compte des relations de cohésion lexicale tout en pouvant être construite automatiquement à partir d'un corpus, ce qui ne restreint pas trop les conditions de son application. F06C est l'extension de F06 qui utilise les relations présentes dans ce réseau pour améliorer la détection des similarités thématiques entre unités de discours en complément de la stricte récurrence lexicale. Nous présenterons également dans cette section le segmenteur F06CT, qui conjugue à la fois l'approche endogène reposant sur les thèmes des documents et l'exploitation d'un réseau de cooccurrences lexicales.

6.1. *Description du réseau de cooccurrences lexicales*

Le réseau de cooccurrences lexicales que nous avons utilisé pour F06C et F06CT concerne le français. Il a été constitué à partir de vingt-quatre mois du journal *Le Monde* sélectionnés entre 1990 et 1994, ce qui représente un corpus d'environ 40 millions de mots. Le corpus initial a d'abord été prétraité afin de caractériser les textes par leurs mots les plus thématiquement significatifs, en l'occurrence les noms, les verbes et les adjectifs, donnés sous forme lemmatisée. Les noms étaient à la fois des noms simples et des noms composés[4]. Les cooccurrences ont ensuite été extraites en utilisant une fenêtre glissante selon la méthode décrite dans (Church et Hanks, 1990). Les paramètres de cette extraction ont été fixés afin de favoriser la capture de relations sémantiques et thématiques : la fenêtre était assez large (20 mots), respectait la fin des textes et l'ordre des cooccurrences n'était pas conservé. Nous avons comme Church et Hanks adopté une estimation de l'information mutuelle en tant que mesure de cohésion de chaque cooccurrence, mesure normalisée dans notre cas par l'information mutuelle maximale relative au corpus. Après filtrage des cooccurrences les moins significatives (cohésion < 0,1 et moins de 10 occurrences), nous avons obtenu un réseau d'approximativement 23 000 mots et 5,2 millions de cooccurrences.

6.2. *Utiliser un réseau de cooccurrences lexicales pour segmenter*

Comme dans le cas de l'utilisation des thèmes du document par F06T, l'utilisation des relations de cooccurrence lexicale par F06C s'effectue lors de l'évaluation de la cohésion au sein de la fenêtre glissante d'analyse. Ces relations, même lorsque leur mesure de cohésion est forte, n'ont pas la fiabilité des relations issues d'une ressource

4. Les noms composés faisaient partie des 2 300 noms composés les plus fréquents observés sur 11 ans du journal *Le Monde*.

sémantique construite manuellement. Nous ne les utilisons donc pas de façon individuelle et directe. Nous évaluons plus exactement à quel degré un mot d'un volet de la fenêtre est lié à l'autre volet de la fenêtre au travers de ces relations de cohésion lexicale. L'évaluation globale de la cohésion au sein de la fenêtre d'analyse s'effectue donc en trois temps :

– le calcul de la cohésion s'appuyant sur la seule récurrence lexicale ;

– la sélection pour chaque volet de la fenêtre de ses mots les plus fortement liés du point de vue de la cohésion lexicale à l'autre volet de la fenêtre ;

– le calcul de la cohésion s'appuyant sur les relations de cooccurrence lexicale et sa combinaison avec la cohésion fondée sur la récurrence lexicale.

La première étape correspond comme pour F06T à l'évaluation de la cohésion faite dans F06. L'étape de sélection des mots de chaque volet de la fenêtre commence quant à elle par un filtrage visant à sélectionner les relations de cooccurrence les plus fiables. Ce filtrage prend la forme d'un seuillage portant à la fois sur leur fréquence et sur leur mesure de cohésion. Ainsi, dans les évaluations présentées à la section 7.3.3, seules les relations dont la fréquence est supérieure ou égale à 14 et la cohésion est supérieure ou égale à 0,14 ont été retenues. Un mot d'un volet de la fenêtre est ensuite sélectionné sur la base du nombre de mots de l'autre volet avec lesquels il est impliqué dans une relation de cooccurrence. Le nombre minimal fixé dans les évaluations de la section 7.3.3 est égal à 2. Le calcul de la cohésion reposant sur les cooccurrences lexicales reprend alors le même principe que pour les segmenteurs précédents. Le nombre de mots sélectionnés dans chaque volet est rapporté au nombre de mots dans la fenêtre. Comme pour F06T, ce calcul exclut les mots déjà impliqués dans une relation de récurrence :

$$cohésion\_cooc(x) = \frac{card(M_{cooc}(g) - M_{réc}) + card(M_{cooc}(d) - M_{réc})}{card(F_g) + card(F_d)} \quad [5]$$

où $M_{cooc}(x)$ représente les mots du volet x de la fenêtre (*g*auche ou *d*roit) sélectionnés sur la base des relations de cooccurrence lexicale et $M_{réc}$, les mots impliqués dans une relation de récurrence.

La cohésion globale de la fenêtre d'analyse est finalement donnée par la somme de $cohésion\_réc(x)$ et de $cohésion\_cooc(x)$.

6.3. *Conjuguer l'utilisation des thèmes et d'un réseau de cooccurrences lexicales*

F06C représente une tentative pour intégrer l'utilisation de ressources externes à une approche endogène, matérialisée en l'occurrence par F06. De façon parallèle, F06CT est une tentative pour intégrer l'utilisation du même type de ressources à une autre approche endogène, incarnée par F06T. L'intégration s'effectue comme précédemment au niveau de l'évaluation de la cohésion au sein de la fenêtre d'ana-

lyse : une valeur de cohésion est calculée selon chaque dimension prise en compte (récurrence lexicale, thèmes du document, relations de cooccurrence lexicale) et ces différentes valeurs sont combinées au travers d'une somme. La cohésion globale calculée pour la fenêtre glissante dans F06CT est donc égale à la somme de $cohésion\_réc(x)$, de $cohésion\_thm(x)$ et de $cohésion\_cooc(x)$ en ne prenant pas en compte dans $cohésion\_cooc(x)$ les mots sélectionnés déjà intégrés dans le calcul de $cohésion\_thm(x)$, autrement dit les mots déjà sélectionnés du fait de leur appartenance à un des thèmes représentatifs de la fenêtre d'analyse.

7. Résultats et évaluation

7.1. *Corpus de travail et d'évaluation*

L'objectif principal de l'évaluation que nous avons menée était de pouvoir comparer entre eux les différents segmenteurs que nous avons présentés ci-dessus et plus spécifiquement de pouvoir juger de l'intérêt d'une approche endogène ou d'une approche exogène pour améliorer un segmenteur thématique fondé sur la récurrence lexicale. L'approche endogène que nous proposons se fonde sur la découverte des thèmes des documents. Comme nous l'avons vu à la section 5.2, une telle découverte s'appuie sur la distribution des mots dans le document. Par conséquent, le cadre d'évaluation proposé par (Choi, 2000), qui est maintenant classiquement utilisé pour l'évaluation des segmenteurs thématiques, n'est pas directement applicable ici. Ce cadre propose en effet de construire artificiellement les documents de référence destinés à évaluer les résultats d'un segmenteur thématique en assemblant des extraits de « vrais » documents. Dans le cas précis du travail rapporté dans (Choi, 2000), chaque document d'évaluation est ainsi constitué de 10 extraits de documents issus du corpus Brown. Chaque extrait, dont la taille est comprise entre 3 et 11 phrases, provient d'un document différent.

Cette procédure présente un double avantage : elle ne nécessite pas l'intervention d'un jugement humain et elle permet de contrôler très précisément les caractéristiques du corpus d'évaluation (taille des segments, des documents, etc.). Son inconvénient principal est évidemment que les documents ainsi constitués sont artificiels et que la tâche évaluée s'apparente plus à une segmentation en documents qu'à une segmentation des documents. Ce caractère artificiel peut avoir une incidence sur l'évaluation relative des systèmes comme l'a montré (Georgescul *et al.*, 2006). Néanmoins, l'utilisation d'un corpus de référence constitué de documents segmentés manuellement ne constitue pas pour autant une alternative très fiable. Même si plusieurs travaux, dont (Passonneau et Litman, 1997), (Klavans *et al.*, 1998) et (Bestgen et Piérard, 2006), s'accordent sur la possibilité de définir une segmentation manuelle de référence fondée sur une corrélation significative des jugements humains, (Bestgen et Piérard, 2006) souligne le fait que les méthodes automatiques de segmentation ont intrinsèquement beaucoup de mal à reproduire ce type de segmentation, probablement parce que les critères utilisés par les annotateurs ne sont pas assez systématiques.

Compte tenu de cette difficulté, nous avons opté pour un cadre d'évaluation proche de celui de Choi dans son principe, notamment du fait de sa simplicité de mise en œuvre, mais prenant en compte le problème soulevé ci-dessus : notre découverte des thèmes d'un document s'appuie sur le fait que les mots d'un thème ont tendance à apparaître conjointement à l'échelle du document. Or, cette hypothèse n'a plus vraiment de sens pour des documents construits suivant les principes proposés par Choi puisque les extraits de documents assemblés sont sans relation thématique. C'est la raison pour laquelle nous avons adapté ces principes pour nous rapprocher d'une forme des documents plus réaliste tout en conservant les avantages de ce cadre d'évaluation.

Cette adaptation concerne la façon dont les extraits de documents agencés sont sélectionnés. Au lieu de tirer chaque extrait d'un document différent, nous n'utilisons que deux documents. Chacun d'entre eux est divisé comme dans le cas de Choi en segments de 3 à 11 phrases et le document d'évaluation est constitué en prenant, à partir du début des documents, alternativement un segment dans un des deux documents et le suivant dans l'autre et ce, jusqu'à obtenir 10 segments ou jusqu'à ce que le processus de construction atteigne la fin d'un des deux documents. Pour nous assurer que deux segments consécutifs font référence à deux thèmes différents et que le changement de thème entre les deux est donc effectif, les deux documents sont sélectionnés de manière à appartenir à des thématiques différentes.

	Français	**Anglais**
Nombre de documents sources	128	87
Nombre de topics sources	11	3
Nombre de segments/document	10 (84 %) 8 (16 %)	10 (97 %) 8 (3 %)
Nombre moyen de phrases/doc.	65	68
Nombre moyen de mots pleins/doc.	797	604

Tableau 1. *Caractéristiques des corpus d'évaluation*

Pour ce faire, nous nous sommes appuyés sur les données constituées pour l'évaluation CLEF, dédiée à la recherche d'information multilingue. Les documents source utilisés pour réaliser notre corpus étaient ainsi des documents issus du corpus CLEF pour lesquels nous disposions d'un jugement de pertinence par rapport à un des topics d'interrogation[5] définis pour l'évaluation, que nous assimilons ici à des thèmes. Chaque document de notre corpus a ainsi été construit à partir de deux documents du corpus CLEF jugés pertinents pour deux topics différents. Outre l'existence de cette forme d'annotation thématique, le corpus CLEF présente l'avantage de comporter des documents comparables pour différentes langues, propriété que nous avons exploitée pour constituer des corpus d'évaluation en français et en anglais assez proches. Plus précisément, ces documents sont des articles de journaux des années 1994 et 1995, *Le*

5. Le terme *topic* est le terme consacré pour désigner la forme spécifique des requêtes d'interrogation utilisées dans les évaluations de recherche d'information telles que TREC ou CLEF.

Monde pour le français, *Los Angeles Times* et *Glasgow Herald* pour l'anglais, et des dépêches de l'agence de presse SDA couvrant la même période dans les deux langues. Le tableau 1 résume les caractéristiques de ces deux corpus. Chacun d'entre eux est formé de 100 documents. On voit donc qu'un document source de CLEF contribue généralement à la construction de plusieurs documents d'évaluation. On constate également que certains d'entre eux peuvent comporter moins de 10 segments lorsque le processus de construction atteint la fin d'un des deux documents avant de parvenir à ce nombre de référence. Il faut préciser que le corpus en anglais n'est destiné, dans le cas présent, qu'à tester la généralité des résultats obtenus par les méthodes endogènes car un réseau de cooccurrences lexicales n'ayant pu être utilisé pour l'anglais, F06C et F06CT n'ont pas pu être testés pour cette langue.

7.2. *Découverte des thèmes d'un document*

Les segmenteurs F06T et F06CT exploitant les thèmes des documents, il est intéressant d'avoir une évaluation de la méthode permettant de les découvrir afin de mettre en relation ses résultats avec ceux de la segmentation thématique. Nous utilisons, pour ce faire, les corpus que nous avons présentés à la section 7.1. Dans ce contexte, la représentation de référence d'un thème d'un document est constituée par le vocabulaire contenu dans les segments relevant de ce thème et n'apparaissant pas dans les segments des autres thèmes. Le tableau 2 donne l'exemple des thèmes découverts pour un document de test construit à partir d'un document à propos des risques induits par le problème de la vache folle en Suisse et d'un document évoquant les difficultés des fabricants de skis suisses. On constate que les deux thématiques sont effectivement bien séparées, mais que certains mots des thèmes découverts (en italique) sont absents des thèmes de référence qui leur correspondent. Un mot comme *Suisse* peut être considéré comme pertinent, mais sa présence dans les deux thèmes explique son absence des thèmes de référence. En revanche, des mots tels que *devenir*, *année* ou *dernier* n'ont pas de spécificité par rapport aux thèmes considérés et ne sont présents dans les thèmes découverts que du fait de leur forte présence dans un des deux documents d'origine.

Thème 1	folle, fédéral, cas, *devenir*, vache, bovin, infecter, maladie, ESB, humain, déclarer
Thème 2	fabricant, Streule, marché, paire, production, ski, Stöckli, *Suisse*, *indiquer*, *directeur*, *année*, entreprise, *dernier*

Tableau 2. *Exemple de thèmes découverts pour un document du corpus d'évaluation*

Pour juger plus formellement de la pertinence des thèmes découverts par rapport aux thèmes de référence, nous faisons appel à trois mesures complémentaires. La principale est la mesure de pureté, classiquement utilisée pour évaluer les résultats d'une classification non supervisée. La pureté d'un thème découvert est donnée par la proportion de son vocabulaire correspondant au vocabulaire du thème de référence auquel

il est associé. Un thème découvert est associé au thème de référence avec lequel il partage le plus de mots. La pureté globale des thèmes découverts est donnée quant à elle par :

$$Pureté = \sum_{i=1}^{k} \frac{v_i}{V} P(Td_i) \quad [6]$$

où $P(Td_i)$ est la pureté du thème découvert Td_i, V est l'ensemble du vocabulaire des thèmes découverts et v_i est le vocabulaire de Td_i. La deuxième mesure que nous utilisons évalue dans quelle mesure les thèmes de référence sont effectivement représentés parmi les thèmes découverts, chacun d'entre eux étant associé comme précédemment au thème de référence partageant avec lui le plus large vocabulaire. Cette mesure est donnée par le rapport entre le nombre de thèmes découverts associés à un thème de référence et le nombre de thèmes de référence. Finalement, la dernière mesure permet de savoir si le vocabulaire des thèmes découverts couvre de façon significative ou pas le vocabulaire des thèmes de référence en calculant le rapport entre la taille du vocabulaire des thèmes découverts faisant partie des thèmes de référence et celle du vocabulaire des thèmes de référence.

	Pureté	**% de thèmes représentés**	**Couverture du vocabulaire des thèmes**
Français	0,771 (0,117)	0,895 (0,239)	0,299 (0,078)
Anglais	0,766 (0,082)	0,990 (0,100)	0,316 (0,053)

Tableau 3. *Résultats de la découverte des thèmes des documents*

Le tableau 3 donne pour chaque mesure sa moyenne, suivie de son écart type. Les résultats sont globalement comparables pour le français et l'anglais même si l'on peut observer une pureté un peu meilleure en français et une représentation des thèmes et de leur vocabulaire un peu meilleure en anglais. À un niveau plus général, on constate que la méthode de découverte des thèmes d'un document produit des représentations de thèmes assez peu bruitées, *i.e.* assez pures, que chaque thème de référence y est habituellement représenté mais que cette représentation ne couvre qu'une petite partie de sa forme de référence. Les thèmes découverts semblent donc pertinents mais leur description est sans doute trop lacunaire.

7.3. *Segmentation thématique d'un document*

7.3.1. *Méthodologie d'évaluation*

Concernant la segmentation thématique, la procédure d'évaluation consiste à appliquer l'algorithme de segmentation considéré sur les documents construits dont on

a supprimé les marques de séparation des segments et à comparer les changements de thème détectés avec ces marques de référence. Cette comparaison s'effectue en utilisant principalement la mesure d'erreur P_k (Beeferman *et al.*, 1999), conformément aux évaluations récentes faites dans ce domaine. P_k évalue la probabilité que deux mots choisis aléatoirement dans un document et séparés par k mots soient jugés comme appartenant au même segment alors qu'ils sont dans des segments différents (faux négatif) ou qu'ils soient jugés comme appartenant à des segments différents alors qu'ils sont dans le même (fausse alarme). k est égal à la moitié de la taille moyenne en mots des segments au niveau du corpus de référence. L'objectif est bien entendu de minimiser P_k. Cette mesure permet en particulier de gérer le problème de la distance entre une borne de segment trouvée et une borne de référence que posent les mesures de précision et de rappel. WindowDiff (Pevzner et Hearst, 2002), dont nous donnons aussi les résultats, est une variante de P_k corrigeant certaines de ses insuffisances en prenant en compte le nombre de frontières de segments séparant deux mots situés dans des segments différents.

7.3.2. *Résultats des approches endogènes*

Le tableau 4 et le tableau 5 donnent non seulement les résultats obtenus par les segmenteur F06 et F06T sur les corpus français et anglais décrits à la section 7.1, mais également les résultats sur ces mêmes corpus de certaines méthodes de référence : U00 est ainsi la méthode décrite dans (Utiyama et Isahara, 2001), C99, celle proposée dans (Choi, 2000) et LCseg est présentée dans (Galley *et al.*, 2003)[6]. TextTiling* est une variante de TextTiling dans laquelle la troisième étape d'identification des changements de thème est reprise de (Galley *et al.*, 2003). Il est à noter que toutes ces méthodes sont utilisées comme F06 et F06T, sans fixer le nombre de changements de thème à trouver et que leurs paramètres ont été adaptés au corpus d'évaluation pour en tirer le meilleur parti. Pour chaque résultat de ces méthodes, nous donnons en outre le degré de signification p de sa différence avec F06 et F06T, niveau évalué grâce à un test de Student unilatéral suivant la pratique initiée par Choi et reprise depuis pour ce type d'évaluation. Ce niveau correspond plus précisément à la probabilité d'erreur quant au fait de rejeter l'hypothèse stipulant que la différence entre les résultats testés n'est pas significative. La probabilité maximale en dessous de laquelle deux résultats sont considérés comme significatifs est classiquement fixée à 0,05. Les différences significatives sont marquées en gras dans les deux tableaux.

Le premier enseignement à tirer de ces deux tableaux de résultats est que l'hypothèse sous-tendant F06T concernant l'intérêt de prendre en compte pour la segmentation les thèmes des documents découverts automatiquement est confirmée. Aussi bien pour le français que pour l'anglais, les résultats de F06T sont significativement et même très significativement supérieurs à ceux de F06. Le deuxième enseignement notable est le fait que ces résultats sont assez stables, même si les deux corpus consi-

6. Pour U00 et LCseg, nous avons repris les implémentations mises à disposition par leurs auteurs. Nous avons réimplémenté C99 en C++ à partir du système originel en Java de Choi en vérifi ant que nous obtenions les mêmes résultats.

Systèmes	P_k (%)			WindowDiff (%)		
	erreur	**p(F06)**	**p(F06T)**	**erreur**	**p(F06)**	**p(F06T)**
U00	25,91	**0,003**	**1,3e-07**	27,42	0,799	**0,032**
C99	27,57	**4,2e-05**	**3,6e-10**	35,42	**8,6e-07**	**6,5e-13**
TextTiling*	21,08	0,699	**0,037**	27,43	0,803	**0,032**
LCseg	20,55	0,439	0,111	28,31	0,767	**0.007**
F06	21,58	/	**0,013**	27,83	/	**0,016**
F06T	*18,46*	**0,013**	/	*24,05*	**0,016**	/

Tableau 4. *Résultats des approches endogènes pour le corpus en français*

dérés sont assez proches : la mise au point de F06 et de F06T a été faite sur le corpus français et les résultats obtenus pour l'anglais sont comparables, aussi bien au niveau de la comparaison entre F06 et F06T qu'au niveau de la comparaison de leurs résultats avec les méthodes de référence.

Concernant cette comparaison, on retrouve au niveau des résultats les parentés entre méthodes : TextTiling*, LCseg, F06 et F06T partagent un nombre important de principes, ce qui se caractérise ici par des résultats significativement plus élevés que ceux de U00 ou C99. Cette tendance vient d'ailleurs partiellement à rebours des résultats obtenus sur le corpus de Choi, ce qui laisse à penser que certaines méthodes performantes sur ce corpus ne le seront pas nécessairement sur de « vrais » documents. Cette observation va aussi dans le sens des constats faits par (Georgescul *et al.*, 2006) et (Bestgen et Piérard, 2006). Les différences relatives entre C99 et TextTiling observées dans (Georgescul *et al.*, 2006) sur le corpus de Choi et un corpus de dialogues sont similaires à celles que nous obtenons, ce qui suggère que nos documents artificiels sont peut-être plus proches de « vrais » documents que ceux de Choi.

Systèmes	P_k (%)			WindowDiff (%)		
	erreur	**p(F06)**	**p(F06T)**	**erreur**	**p(F06)**	**p(F06T)**
U00	19,42	**0,048**	**4,3e-05**	21,22	0,826	**0,039**
C99	21,63	**1,2e-04**	**1,8e-09**	30,64	**1,4e-12**	**0**
TextTiling*	15,81	0,308	0,111	19,80	0,355	0,253
LCseg	14,78	**0,043**	0,496	19,73	0,325	0,271
F06	16,90	/	**0,010**	20,93	/	**0,046**
F06T	*14,06*	**0,010**	/	*18,31*	**0,046**	/

Tableau 5. *Résultats des approches endogènes pour le corpus en anglais*

Enfin, on pourra noter, sans véritablement l'expliquer, que les performances de toutes ces méthodes sont plus élevées en anglais qu'en français. Compte tenu de la similarité des deux corpus, la différence semble en première analyse imputable à la

langue, peut-être du fait de la moindre réticence stylistique à la répétition en anglais. Mais de plus amples analyses seraient nécessaires pour l'affirmer véritablement.

7.3.3. *Résultats des approches exogènes*

Le tableau 6 donne les résultats de F06C et de F06CT sur le corpus d'évaluation français tandis que le tableau 7 fournit le degré de signification de ces résultats par rapport aux résultats de toutes les autres méthodes. Le premier constat est que l'utilisation de relations de cooccurrences lexicales induit comme l'utilisation de thèmes une amélioration significative des résultats de segmentation. Les résultats de F06C sont ainsi significativement plus élevés que ceux de F06. Les résultats de F06C sont aussi significativement plus élevés que ceux de toutes les méthodes de référence que nous avons appliquées, ce qui situe sur ce plan F06C un peu plus haut que F06T.

Systèmes	$\mathbf{P_k}$ **(%)**	**WindowDiff (%)**
F06C	16,48	20,94
F06CT	14,59	18,41

Tableau 6. *Résultats des approches exogènes pour le corpus en français*

Le tableau 7 conduit néanmoins à nuancer une première impression accordant un avantage incontestable à l'utilisation de connaissances externes par rapport à une approche endogène. Il montre en effet que la différence entre les résultats de F06T et ceux de F06C n'est pas significative pour P_k et que dans le cas de WindowDiff, le degré de signification n'est pas parmi les plus faibles, bien que suffisant pour valider la différence. Par conséquent, même si les résultats de F06C sont globalement supérieurs à ceux de F06T, F06T rivalise de façon assez équilibrée avec F06C. Ce constat est d'autant plus intéressant que F06T ne fait appel à aucune connaissance extérieure au document, ce qui le rend d'usage beaucoup plus large.

Systèmes	**p(F06C)**		**p(F06CT)**	
	$\mathbf{P_k}$	**WindowDiff**	$\mathbf{P_k}$	**WindowDiff**
U00	**8,7e-11**	**3,6e-05**	**1,8e-14**	**3,6e-09**
C99	**1,7e-12**	**0**	**1,6e-12**	**0**
TextTiling*	**3,2e-04**	**3,5e-05**	**1,8e-07**	**3,5e-09**
LCseg	**0,002**	**1,6e-06**	**3,3e-06**	**2,1e-10**
F06	**6,5e-05**	**4,8e-06**	**2,6e-08**	**7,3e-10**
F06T	0,107	**0,037**	**0,001**	**1,2e-04**
F06C	⁄	⁄	0,111	0,072

Tableau 7. *Degré de signifi cation des résultats des approches exogènes*

Les tableaux 6 et 7 montrent aussi la complémentarité de l'utilisation d'une approche endogène et d'une approche exogène au travers des résultats de F06CT. Ce

segmenteur obtient en effet les meilleurs résultats parmi tous les segmenteurs testés, aussi bien en ce qui concerne le niveau absolu de ses performances que leur degré de signification. F06CT est de fait significativement plus performant que F06T pour toutes les mesures, ce que F06C ne parvient pas à faire. Ce phénomène est sans doute une manifestation de la nécessaire complémentarité des deux approches. Quelle que soit son extension, une source de connaissances ne peut être exhaustive. Par exemple, elle ne peut englober tous les noms propres susceptibles d'apparaître dans un texte. À l'inverse, découvrir les rapprochements entre les mots d'un document en s'appuyant sur leur seule distribution dans le document n'est pas suffisant pour extraire toutes les relations sémantiques existant entre ces mots. Il est donc probable qu'une source de connaissances externe soit utile pour mettre en évidence les relations sémantiques existant entre les termes les plus usités tandis qu'une approche endogène est utile pour trouver ces relations entre des termes plus spécifiques du document traité.

Pour finir, il faut noter que F06CT ne parvient pas à se différencier significativement de F06C, ce que l'on peut interpréter par un poids plus fort pris par les connaissances externes par rapport aux relations endogènes.

8. Conclusion et perspectives

Dans cet article, nous avons présenté une méthode de segmentation thématique fondée sur la récurrence lexicale inspirée de *TextTiling* et nous avons proposé deux approches pour en améliorer les résultats. Chacune d'elles a pour objectif d'assurer une meilleure détection des similarités thématiques entre les unités de discours composant les segments. La première approche réalise cette amélioration de manière endogène en exploitant, pour ce faire, les thèmes du document à segmenter qu'elle découvre de façon non supervisée. La seconde approche est exogène et utilise pour le même but des relations sélectionnées dans un réseau de cooccurrences lexicales constitué automatiquement à partir d'un large corpus. Un segmenteur combinant ces deux approches a en outre été testé. Pour évaluer ces segmenteurs thématiques, nous avons proposé une adaptation du cadre d'évaluation proposé par Choi afin de se rapprocher de documents plus réalistes tout en ne sacrifiant pas la simplicité de mise en œuvre qui le caractérise. Les évaluations menées sur un corpus en français et en anglais pour l'approche endogène et en français seulement pour l'approche exogène ont globalement montré que la première approche rivalise de façon équilibrée avec la seconde. L'approche exogène s'avère légèrement meilleure, mais demande des moyens plus importants pour son application. Par ailleurs, ces évaluations ont également montré qu'approche endogène et approche exogène sont complémentaires et que les combiner s'avère intéressant.

La dernière phase du travail que nous avons présenté a été de combiner pour la segmentation thématique une approche endogène et une approche exogène. Une de ses extensions les plus directes serait d'appliquer cette même combinaison à la découverte de thèmes, c'est-à-dire à l'identification thématique. Ainsi, le réseau de cooccurrences lexicales utilisé pour la segmentation pourrait aussi servir de source de connaissances pour identifier les thèmes d'un document. L'algorithme de découverte

des thèmes travaillant à partir d'un graphe de similarité, son application à un réseau de cooccurrences lexicales apparaît assez directe. Une attention toute particulière devra en revanche être accordée à la combinaison des relations de cooccurrence issues du réseau et de celles issues du document traité afin de tirer avantage des complémentarités entre ces deux sources de relations.

La seconde principale extension que nous envisageons pour ce travail concerne la nature de la source de connaissances externe utilisée. Bien que la cohésion lexicale portée par un réseau de cooccurrences lexicales constitue un indicateur intéressant de cohésion thématique, il peut être intéressant d'utiliser une source de connaissances plus spécifiquement thématique. Le réseau lexical thématique proposé dans (Ferret et Zock, 2006) est une possibilité allant dans ce sens que nous souhaitons expérimenter.

Enfin, en se situant dans un contexte plus large, il serait nécessaire d'évaluer l'impact des améliorations constatées au niveau de la segmentation dirigée par le contenu sur un système plus global tel que celui décrit dans (Couto *et al.*, 2004), associant ce type de segmentation avec une méthode exploitant des marques discursives. Au-delà, certaines des marques discursives utilisées dans (Couto *et al.*, 2004), en l'occurrence les introducteurs de cadres thématiques, permettent de réaliser une forme d'identification des thèmes d'un document[7]. Il serait donc intéressant d'associer les résultats de cette identification thématique à ceux de la découverte des thèmes présentée dans cet article afin de créer une synergie entre ces deux approches, comme c'est le cas au niveau de la segmentation.

Remerciements

Nous remercions Michel Galley pour avoir mis à notre disposition l'implémentation de son système LCseg.

9. Bibliographie

Barzilay R., Elhadad M., « Using Lexical Chains For Text Summarization », *ACL 97 Workshop on Intelligent Scalable Text Summarization*, Madrid, Spain, p. 10-17, 1997.

Beeferman D., Berger A., Lafferty J., « Statistical Models for Text Segmentation », *Machine Learning*, vol. 34, n° 1, p. 177-210, 1999.

Bestgen Y., Piérard S., « Comment évaluer les algorithmes de segmentation automatique? Essai de construction d'un matériel de référence. », *TALN 2006*, Leuven, Belgium, p. 407-414, 2006.

Bigi B., Mori R. d., El-Bèze M., Spriet T., « Detecting topic shifts using a cache memory », *5^{th} International Conference on Spoken Language Processing*, p. 2331-2334, 1998.

7. Dans une phrase commençant par *En ce qui concerne le problème du fret ferroviaire ...*, l'introducteur de cadre *en ce qui concerne* est une marque de segmentation et le groupe nominal qui la suit caractérise le thème du nouveau segment.

Boguraev B., Neff M. S., « Discourse Segmentation in Aid of Document Summarization », *HICSS*, 2000.

Caillet M., Pessiot J.-F., Amini M., Gallinari P., « Unsupervised Learning with Term Clustering for Thematic Segmentation of Texts », *7th Conference on Rercherche d'Information Assistée par Ordinateur (RIAO'04)*, p. 1-11, 2004.

Charolles M., « L'encadrement du dicours : Univers, Champs, Domaines et Espaces », *Cahier de Recherche Linguistique*, vol. 6, p. 1-73, 1997.

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

Choi F. Y. Y., Wiemer-Hastings P., Moore J., « Latent Semantic Analysis for Text Segmentation », *EMNLP'01*, p. 109-117, 2001.

Church K. W., Hanks P., « Word Association Norms, Mutual Information, And Lexicography », *Computational Linguistics*, vol. 16, n° 1, p. 22-29, 1990.

Châar S. L., Ferret O., Fluhr C., « Filtrage pour la construction de résumés multi-documents guidée par un profi l », *Traitement Automatique des Langues*, vol. 45, n° 1, p. 65-93, 2004.

Couto J., Ferret O., Grau B., Hernandez N., Jackiewicz A., Minel J.-L., Porhiel S., « RÉGAL, un système pour la visualisation sélective de documents », *Revue d'Intelligence Artificielle*, vol. 18, n° 4, p. 481-514, 2004.

Dias G., Alves E., « Discovering Topic Boundaries for Text Summarization based on Word Co-occurrence », *Recent Advances in Natural Language Processing (RANLP 2005)*, Borovets, Bulgaria, p. 187-191, 2005.

Ertöz L., Steinbach M., Kuma V., « Finding Topics in Collections of Documents: A Shared Nearest Neighbor Approach », *Text Mine'01, Workshop of the 1st SIAM International Conference on Data Mining*, 2001.

Fellbaum C. (ed.), *WordNet:An Electronic Lexical Database*, The MIT Press, 1998.

Ferret O., « How to thematically segment texts by using lexical cohesion? », *ACL-COLING'98*, p. 1481-1483, 1998.

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

Ferret O., « Découvrir les thèmes d'un document pour en améliorer la segmentation thématique », *Conférence Internationale sur le Document Electronique (CIDE 9)*, p. 97-111, 2006.

Ferret O., Grau B., « A Topic Segmentation of Texts based on Semantic Domains », *ECAI 2000*, Berlin, Germany, p. 426-430, 2000.

Ferret O., Grau B., Masson N., « Thematic segmentation of texts: two methods for two kinds of texts », *ACL-COLING'98*, vol. 1, Montréal, Canada, p. 392-396, 1998.

Ferret O., Zock M., « Enhancing Electronic Dictionaries with an Index Based on Associations », *COLING-ACL 2006*, Sydney, Australia, p. 281-288, 2006.

Galley M., McKeown K., Fosler-Lussier E., Jing H., « Discourse Segmentation of Multi-party Conversation », *41st Annual Meeting of the Association for Computational Linguistics (ACL-03)*, p. 562-569, 2003.

Georgescul M., Clark A., Armstrong S., « An Analysis of Quantitative Aspects in the Evaluation of Thematic Segmentation Algorithms », *7th SIGdial Workshop on Discourse and Dialogue*, Sydney, Australia, p. 144-151, 2006.

Hearst M. A., « Multi-paragraph segmentation of expository text », *32th Annual Meeting of the Association for Computational Linguistics*, p. 9-16, 1994.

Jobbins A. C., Evett L. J., « Text Segmentation Using Reiteration and Collocation », *ACL-COLING'98*, p. 614-618, 1998.

Kan M.-Y., Klavans J. L., McKeown K. R., « Linear Segmentation and Segment Relevence », *Proceedings of 6th International Workshop of Very Large Corpora (WVLC-6)*, p. 197-205, 1998.

Kaufmann S., « Cohesion and Collocation: Using Context Vectors in Text Segmentation », *37th Annual Meeting of the Association for Computational Linguistics (Student Session)*, p. 591-595, 1999.

Klavans J. L., McKeown K. R., Kan M.-Y., Lee S., « Resources for Evaluation of Summarization Techniques », *1st International Conference on Language Resources and Evaluation (LREC'98)*, Granada, Spain, 1998.

Kozima H., « Text Segmentation Based on Similarity between Words », *31th Annual Meeting of the Association for Computational Linguistics (Student Session)*, p. 286-288, 1993.

Landauer T., Foltz P., Laham D., « An introduction to Latent Semantic Analysis », *Discourse Processes*, vol. 25, p. 259-284, 1998.

Morris J., Hirst G., « Lexical Cohesion Computed by Thesaural Relations as an Indicator of the Structure of Text », *Computational Linguistics*, vol. 17, n° 1, p. 21-48, 1991.

Passonneau R. J., Litman D. J., « Discourse Segmentation by Human and Automated Means », *Computational Linguistics*, vol. 23, n° 1, p. 103-139, 1997.

Pelleg D., Moore A., « X-means: Extending K-means with Effi cient Estimation of the Number of Clusters », *17th International Conference on Machine Learning (ICML)*, p. 727-734, 2000.

Pevzner L., Hearst M. A., « A critique and improvement of an evaluation metric for text segmentation », *Computational Linguistics*, vol. 28, n° 1, p. 19-36, 2002.

Ponte J. M., Croft B. W., « Text segmentation by topic », *First European Conference on research and advanced technology for digital libraries*, 1997.

Salton G., McGill M. J., *Introduction to modern information retrieval*, McGraw Hill, New York, 1983.

Schmid H., « Probabilistic Part-of-Speech Tagging Using Decision Trees », *International Conference on New Methods in Language Processing*, 1994.

Stokes N., Carthy J., Smeaton A., « Segmenting Broadcast News Streams using Lexical Chains », *STarting AI Researchers Symposium, (STAIRS 2002)*, p. 145-154, 2002.

Tür G., Tür D. H., Stolcke A., Shriberg E., « Integrating Prosodic and Lexical Cues for Automatic Topic Segmentation », *Computational Linguistics*, vol. 27, n° 1, p. 31-57, 2001.

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

Xu J., Croft W., « Query Expansion Using Local and Global Document Analysis », *19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Zurich, Switzerland, p. 4-11, 1996.

Yamron J., Carp I., Gillick L., Lowe S., van Mulbregt P., « A hidden Markov model approach to text segmentation and event tracking », *IEEE Conference on Acoustics, Speech and Signal Processing*, p. 333-336, 1998.