Segmenter et structurer thématiquement des textes par l'utilisation conjointe de collocations et de la récurrence lexicale

Olivier Ferret

CEA – LIST
92265 Fontenay-aux-Roses Cedex
olivier.ferret@cea.fr

Résumé – Abstract

Nous exposons dans cet article une méthode réalisant de façon intégrée deux tâches de l'analyse thématique : la segmentation et la détection de liens thématiques. Cette méthode exploite conjointement la récurrence des mots dans les textes et les liens issus d'un réseau de collocations afin de compenser les faiblesses respectives des deux approches. Nous présentons son évaluation concernant la segmentation sur un corpus en français et un corpus en anglais et nous proposons une mesure d'évaluation spécifiquement adaptée à ce type de systèmes.

We present in this paper a method for achieving in an integrated way two tasks of topic analysis: segmentation and link detection. This method combines the lexical recurrence in texts and the relations from a collocation network to compensate for the respective weaknesses of the two approaches. We report its evaluation for segmentation on a corpus in French and another in English and we propose an evaluation measure that specifically suits that kind of systems.

Mots Clés – Keywords

Analyse du discours, analyse thématique, segmentation, détection de liens thématiques.
Discourse analysis, topic analysis, topic segmentation, link detection.

1 Introduction

L'analyse thématique, dont le but est d'identifier les thèmes d'un texte[1], d'en délimiter l'extension et de trouver les relations entre les segments ainsi délimités, a fait l'objet d'un nombre important de travaux récents. La plupart d'entre eux sont consacrés à la segmentation

[1] Nous définissons ici le thème comme une configuration d'unités sémantiques de nature inter-textuelle, c'est-à-dire observable entre plusieurs textes et plusieurs auteurs.

thématique ou s'inscrivent dans le cadre de l'évaluation TDT (Topic Detection and Tracking). Cette dernière aborde les trois axes évoqués ci-dessus mais en se plaçant dans des domaines restreints et en les considérant principalement sous l'angle de tâches indépendantes.

Les systèmes implémentant ces travaux peuvent être catégorisés en fonction du type des connaissances dont ils font usage. La plus grande partie de ceux dédiés à la segmentation thématique, c'est-à-dire au découpage des textes en segments thématiquement homogènes, s'appuient uniquement sur les caractéristiques intrinsèques des textes tels que la distribution des mots (Hearst, 1997 ; Choi, 2000 ; Utiyama, Isahara, 2001) ou des marqueurs linguistiques (Passonneau, Litman, 1997). Ils peuvent être utilisés sans restriction quant au domaine abordé mais leurs performances sont faibles lorsque la structure thématique des textes ne transparaît pas au travers des marques de surface qu'ils exploitent. Un second type de systèmes utilise des connaissances caractérisant la notion de cohésion lexicale. Ces connaissances, non liées à un domaine, prennent la forme d'un réseau de mots construit à partir d'un dictionnaire (Kozima, 1993 ; Morris, Hirst, 1991) ou d'un large ensemble de collocations issues d'un corpus (Ferret, 1998 ; Kaufmann, 1999 ; Choi, 2001). Grâce aux relations entre mots qu'elles contiennent (synonymie, hyperonymie …), ces connaissances permettent d'écarter des changements de thème erronés définis sur la base de ruptures existant au niveau de la récurrence des mots. Un dernier type de systèmes s'appuie sur des connaissances directement liées aux thèmes apparaissant dans les textes qu'ils traitent. Dans le cas de TDT par exemple, ces connaissances sont construites de manière automatique à partir d'un ensemble de textes de référence caractérisant chaque thème considéré. (Bigi *et al.*, 1998) se situe dans la même perspective mais s'intéresse à des thèmes plus larges que les événements propres à TDT. Ces systèmes ont un champ d'action assez limité du fait de leur dépendance vis-à-vis de représentations de thèmes. À l'intérieur de ce champ d'action, elles leur permettent en revanche d'être plus précis.

La méthode d'analyse thématique que nous proposons dans cet article se distingue des travaux que nous venons d'évoquer sur deux points principaux. Tout d'abord, elle vise à assurer simultanément une segmentation thématique des textes et une mise en évidence des liens entre les segments faisant référence au même thème (cf. tâche Link Detection de TDT), première étape vers la structuration thématique des textes. Ensuite, elle met en œuvre une approche hybride : elle s'appuie en effet sur une ressource rendant compte de la cohésion lexicale, en l'occurrence un réseau de collocations, mais l'exploite en conjonction avec la récurrence lexicale dans les textes. Nous détaillons dans cet article l'implémentation de cette méthode par le système TOPICOLL, son évaluation en matière de segmentation pour le français et l'anglais ainsi qu'une évaluation plus globale de ses capacités par une nouvelle mesure.

2 Vue d'ensemble du système TOPICOLL

Dans le prolongement de nombreux travaux portant sur la segmentation du discours, TOPICOLL traite les textes linéairement : il détecte les changements de thème et identifie les liens entre segments sans différer sa décision, c'est-à-dire en tenant compte uniquement des éléments qu'il a pu extraire de la partie du texte déjà analysée. Une fenêtre délimitant l'espace de focalisation de l'analyse est déplacée sur l'ensemble du texte considéré. Cette fenêtre contient sous forme lemmatisée les mots pleins du texte issus de son pré-traitement. Un contexte thématique est associé à cette fenêtre de focalisation. Il est constitué à la fois des mots de la fenêtre et des mots d'un réseau de collocations jugés les plus fortement liés aux

mots de la fenêtre. Les segments se voient également associer un contexte thématique. Celui-ci résulte de la fusion des contextes associés à la fenêtre de focalisation lorsque celle-ci se situe dans leur espace. Un changement de thème est détecté dès lors qu'une différence significative et durable est observée entre le contexte de la fenêtre et celui du segment dans laquelle elle se trouve. La détection de liens thématiques s'effectue quant à elle en comparant le contexte thématique de chaque nouveau segment avec celui des segments précédemment définis. TOPICOLL reprend donc le principe général d'analyse du système SEGAPSITH (Ferret, Grau, 2000) en introduisant au niveau des contextes thématiques la caractérisation de la cohésion lexicale développée dans le cadre du système SEGCOHLEX (Ferret, 1998).

L'utilisation d'un réseau de collocations[2] permet à TOPICOLL de trouver des relations entre les mots au-delà de la simple réitération et d'associer à chaque segment une représentation thématique plus riche, ce qui facilite des tâches comme la détection de liens thématiques. Néanmoins, des travaux tels que (Kozima, 1993), (Ferret, 1998) ou (Kaufmann, 1999) ont montré que le recours à des connaissances lexicales générales n'améliore souvent pas les performances par rapport à l'exploitation de la seule distribution des mots dans les textes. Les méthodes utilisées ne contrôlent pas en effet assez précisément le type des relations qu'elles sélectionnent et ne tiennent pas compte de l'incomplétude de leurs connaissances. De ce fait, en même temps qu'elles permettent d'invalider à juste titre certains des changements de thème correspondant à une rupture dans l'usage du vocabulaire, elles introduisent des ruptures thématiques incorrectes du fait de l'absence dans leurs connaissances des relations lexicales adéquates ou passent à côté de changements de thème réels à cause de la sélection de relations lexicales non pertinentes du point de vue thématique. En combinant la récurrence des mots et la sélection de relations dans un réseau de collocations, TOPICOLL vise donc à exploiter de façon plus précise une source de connaissances non liée à un domaine.

3 Les réseaux de collocations

TOPICOLL s'appuie sur un réseau de collocations propre à chaque langue qu'il traite. Celui pour le français a été construit à partir de 24 mois du journal *Le Monde* sélectionnés entre 1990 et 1994 en respectant un équilibre entre les années et entre les mois ; celui pour l'anglais à partir de deux ans du journal *Los Angeles Times*, issus du corpus TREC. Dans les deux cas, le corpus contient environ 40 millions de mots. Pour chaque réseau, le corpus initial a d'abord été pré-traité 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 ambiguïtés de lemmatisation ont été levées grâce à un étiqueteur morpho-syntaxique. Les collocations ont ensuite été extraites en utilisant une fenêtre glissante selon la méthode décrite dans (Church, Hanks, 1990). Les paramètres de cette extraction ont été fixés pour favoriser la capture de relations thématiques : une fenêtre assez large (20 mots), respectant la fin des textes et ne conservant pas l'ordre des collocations. Nous avons comme Church et Hanks adopté une évaluation de l'information mutuelle en tant que mesure de cohésion des collocations, mesure normalisée dans notre cas par l'information mutuelle maximale relative au corpus. Après filtrage des collocations les moins significatives (cohésion < 0,1 et moins de 10 occurrences), nous avons obtenu un réseau d'environ 23.000

[2] Les nœuds de ce réseau sont constitués par des mots et les arêtes, par des collocations entre ces mots.

mots et 5,2 millions de collocations pour le français et un réseau de 30.000 mots et 4,8 millions de collocations pour l'anglais.

4 Description du système TOPICOLL

TOPICOLL fonde son fonctionnement sur la création, la mise à jour et l'utilisation d'une représentation thématique des segments qu'il définit et du contenu de sa fenêtre de focalisation à chaque position d'un texte. Ces représentations sont appelées des *contextes thématiques*.

4.1 Les contextes thématiques

Un contexte thématique caractérise la dimension thématique de l'entité à laquelle il est associé par l'intermédiaire de deux vecteurs : le *vecteur texte* et le *vecteur collocation*. Chacune de leurs coordonnées représente un mot et sa valeur correspond à un poids traduisant l'importance relative de ce mot par rapport aux autres mots du vecteur. Le *vecteur texte* est composé des mots venant du texte analysé tandis que le *vecteur collocation* contient les mots du réseau de collocations considérés comme fortement liés aux mots du texte.

4.1.1 Le contexte thématique de la fenêtre de focalisation (Cf)

Le *vecteur texte* du contexte associé à la fenêtre de focalisation est constitué des mots pleins de cette fenêtre. Leur poids combine leur importance dans la partie du texte délimitée par la fenêtre, donnée par leur nombre d'occurrences, et leur degré de spécificité hors contexte, exprimé comme dans (Kozima, 1993) par l'information normalisée par rapport à un corpus de référence, en l'occurrence celui ayant permis la construction du réseau de collocations utilisé.

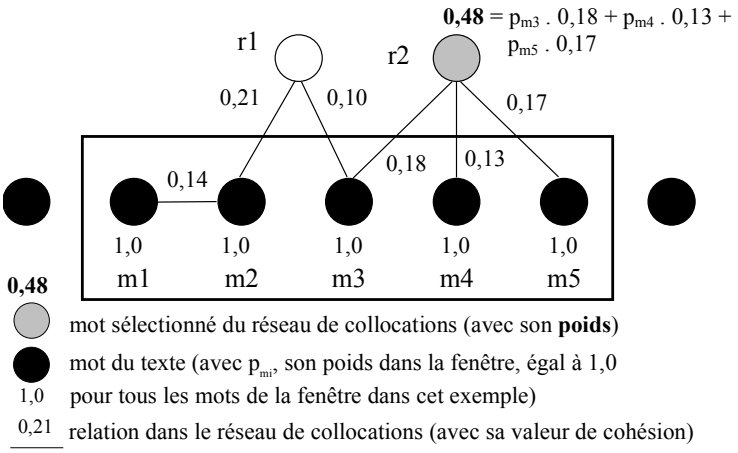

Figure 1 : Sélection et pondération des mots du réseau de collocations

La construction du *vecteur collocation* du contexte de la fenêtre de focalisation s'inspire de la procédure présentée dans (Ferret, 1998) pour évaluer la cohésion lexicale d'un texte. Elle consiste à sélectionner les mots du réseau de collocations thématiquement proches de ceux de la fenêtre. Nous faisons l'hypothèse que cette proximité est liée au nombre de liens existant entre un mot du réseau et les mots de la fenêtre. Un mot du réseau est ainsi retenu s'il est lié à un nombre minimum – 3 dans nos expériences – de mots de la fenêtre. Chaque mot retenu se

voit ensuite assigner un poids, égal à la somme des contributions des mots de la fenêtre auxquels il est lié (cf. Figure 1). La contribution d'un mot de la fenêtre est le résultat de la combinaison, selon une moyenne géométrique, de son poids dans la fenêtre et de la valeur de cohésion entre les deux mots dans le réseau. On obtient donc pour une position *t* de la fenêtre :

$$poids_{coll}(m, Cf, t) = \sum_i \sqrt{poids_{txt}(m_i, Cf, t) \cdot coh(m, m_i)} \quad (1)$$

avec $coh(m,m_i)$, la valeur de cohésion entre *m* et m_i dans le réseau de collocations.

4.1.2 Le contexte thématique d'un segment (Cs)

Le contexte thématique d'un segment est le produit de la fusion des contextes associés à la fenêtre de focalisation lorsque celle-ci se trouvait dans le segment. Cette fusion est réalisée à chaque nouvelle extension du segment : le contexte associé à la nouvelle position englobée est aussitôt combiné au contexte courant du segment. Cette combinaison, réalisée séparément pour les vecteurs texte et collocation, consiste à fusionner deux listes de mots pondérés : les mots du contexte de la fenêtre qui ne sont pas dans celui du segment y sont ajoutés ; le poids des mots de la liste résultante est calculé en fonction de leur poids dans le contexte de la fenêtre et de leur précédent poids dans le contexte du segment :

$$poids_x(m, Cs, t) = poids_x(m, Cs, t-1) + (signif(m) \cdot poids_x(m, Cf, t)) \quad (2)$$

avec *Cf*, le contexte de la fenêtre, *Cs*, celui du segment, $poids_x(m, C_{\{s,f\}}, t)$, le poids du mot *m* dans le vecteur *x* (*txt* ou *coll*) du contexte $C_{\{s,f\}}$ pour la position *t* et *signif(m)*, le degré de spécificité hors contexte de *m* (cf. 4.1.1). Pour les mots de *Cf* initialement absents de *Cs*, $poids_x(m, C_s, t-1)$ est égal à 0. Pour ceux de *Cs*, la réévaluation du poids donnée par (2) représente un compromis concernant la vitesse d'évolution des contextes de segment : elle permet de s'affranchir des micro-variations dans la façon dont un thème est exprimé au sein d'un segment tout en suivant les évolutions thématiques du contexte de la fenêtre de suffisamment près pour ne pas passer à côté d'un changement de thème.

4.1.3 La similarité entre contextes thématiques

Pour déterminer si le contenu de la fenêtre de focalisation est thématiquement cohérent avec le segment courant, on compare les contextes associés à ces deux entités. Cette comparaison est réalisée en deux étapes : une mesure de similarité est d'abord calculée entre les vecteurs des deux contextes ; les valeurs obtenues sont ensuite exploitées par une procédure de décision statuant sur la similarité des deux contextes. De même que (Choi, 2000) et (Kaufmann, 1999), nous utilisons la mesure du cosinus pour évaluer le degré de similarité entre un vecteur du contexte de la fenêtre (*Vf*) et le vecteur de même type dans le contexte du segment (*Vs*) :

$$sim(Vs_x, Vf_x, t) = \frac{\sum_i poids_x(m_i, Cs, t) \cdot poids_x(m_i, Cf, t)}{\sqrt{\sum_i poids_x(m_i, Cs, t)^2 \cdot \sum_i poids_x(m_i, Cf, t)^2}} \quad (3)$$

où $poids_x(m_i, C_{\{s,f\}}, t)$ est le poids du mot m_i dans le vecteur *x* (*txt* ou *coll*) du contexte $C_{\{s,f\}}$.

Afin de minimiser le bruit dans les vecteurs, cette mesure ne prend en compte que les mots les plus récurrents des contextes des segments, l'importance d'un mot dans un contexte étant supposée corrélée avec sa récurrence au sein de celui-ci. Cette récurrence est définie comme la proportion, parmi les contextes de fenêtre de focalisation ayant permis de constituer le contexte de segment, de ceux contenant le mot considéré. Ce rapport doit être supérieur à un seuil fixé pour que le mot soit retenu. Du fait de la plus grande hétérogénéité des mots venant des textes, ce seuil est plus exigeant pour les vecteurs texte que pour les vecteurs collocation.

La procédure de décision s'inspire quant à elle des travaux relatifs à la combinaison de plusieurs systèmes réalisant la même tâche. Dans le cas présent, l'évaluation de la similarité entre les contextes *Cs* et *Cf* s'appuie sur un vote synthétisant le point de vue des vecteurs texte et celui des vecteurs collocation. La valeur de similarité obtenue grâce à (3) est d'abord discrétisée pour chaque type de vecteurs par comparaison avec un seuil fixe : un vote positif en faveur de la similarité des contextes est décidé si la valeur est supérieure à ce seuil. Au final, la similarité des contextes n'est rejetée que si le vote des deux types de vecteurs est négatif.

4.2 Segmentation thématique

L'algorithme de segmentation thématique de TOPICOLL reprend dans son principe celui présenté dans (Ferret, Grau, 2000) : si la similarité entre le contexte de la fenêtre de focalisation et celui du segment actif est rejetée (cf. 4.1.3), TOPICOLL en déduit la présence d'un changement de thème à la position correspondante et le segment actif est clos. Sinon, le segment actif est étendu afin d'englober la position courante. Cet algorithme de base suppose que la phase de transition entre deux segments soit ponctuelle et sans ambiguïté. En réalité, la similarité entre contextes peut être localement fluctuante du fait de la forme de surface des textes. Il est donc préférable d'introduire un délai avant de décider véritablement si le segment actif se termine ou si un nouveau segment s'ouvre. Pour tenir compte de cette incertitude, l'algorithme de segmentation prend la forme d'un automate (cf. Figure 2) dont les transitions sont contrôlées par les trois paramètres suivants :

- l'état courant de TOPICOLL ;
- la similarité entre le contexte de la fenêtre de focalisation et le contexte du segment courant : *Sim* ou *non Sim* ;
- le nombre de positions successives de la fenêtre de focalisation caractérisées par un même état courant de TOPICOLL : *nbConfirm*. Il doit être supérieur à $S_{confirm}$ pour sortir des états *DétectionNouveauThème* et *DétectionChangementThème*.

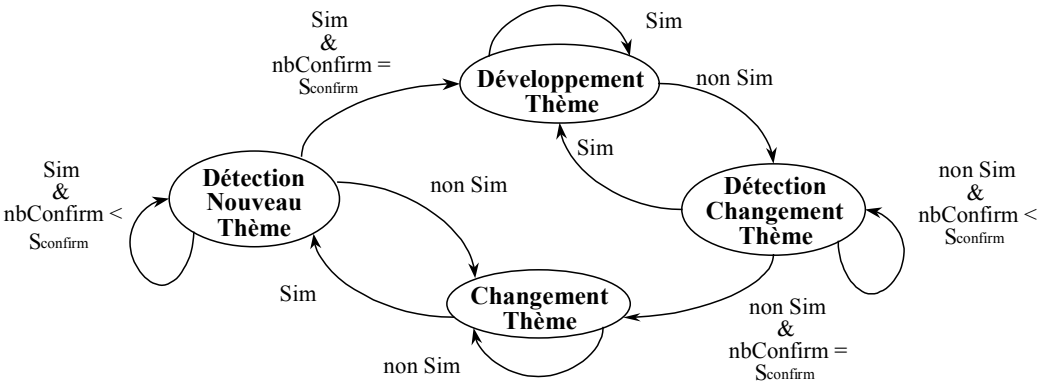

Figure 2 : L'automate de détection des changements de thème

Au début d'un nouveau texte ou à la suite de la détection de la fin d'un segment, TOPICOLL se trouve dans l'état *ChangementThème*. Dès que le contexte thématique de la fenêtre de focalisation demeure stable d'une position à une autre au regard de (3), il entre dans l'état *DétectionNouveauThème*. Il ne peut ensuite atteindre l'état *DéveloppementThème* que si cette stabilité est conservée pour les *nbConfirm - 1* positions suivantes. Sinon, il fait l'hypothèse qu'il s'agit d'une fausse alarme et revient à l'état *ChangementThème*. La détection de la fin d'un segment est le symétrique de la détection de son début. TOPICOLL entre dans l'état *DétectionChangementThème* dès que le contexte de la fenêtre de focalisation change de façon significative entre deux positions successives. La transition vers l'état *ChangementThème* n'est cependant opérée que si ce changement se confirme pour les *nbConfirm - 1* positions suivantes.

4.3 Détection de liens thématiques

L'algorithme de TOPICOLL pour détecter les liens d'identité thématique entre segments est étroitement lié à son algorithme de segmentation. Avant d'entrer dans l'état *DéveloppementThème*, TOPICOLL vérifie si le contexte du nouveau segment est similaire au contexte de l'un des segments déjà définis. La similarité repose dans ce cas sur l'application de (3) entre les vecteurs collocation des contextes et la comparaison par rapport à un seuil spécifique. Si l'une des valeurs de similarité dépasse ce seuil, le nouveau segment est lié au segment correspondant et adopte son contexte comme contexte propre. TOPICOLL fait ainsi l'hypothèse que le nouveau segment continue à développer le thème déjà abordé. Lorsque plusieurs segments sont possibles, TOPICOLL choisit celui pour lequel la similarité des contextes est la plus forte.

5 Expérimentations

5.1 Segmentation thématique

Pour évaluer les capacités de segmentation thématique de TOPICOLL, nous l'avons appliqué à la tâche classique de redécouverte des frontières d'un ensemble de textes concaténés. Pour la mesure des performances, nous avons utilisé la mesure d'erreur probabiliste P_k proposée dans

(Beeferman *et al.*, 1999) et maintenant largement utilisée[3]. Nous avons également calculé la précision et le rappel afin de permettre la comparaison avec certains systèmes plus anciens.

5.1.1 Évaluation de la segmentation pour le français : corpus du journal Le Monde

L'évaluation pour le français a été réalisée sur un ensemble de 49 textes longs de 133 mots en moyenne, extraits du journal *Le Monde* (1995) et couvrant 11 thèmes. Les résultats du Tableau 1 sont des moyennes obtenues sur 10 ordonnancements différents de ces textes. Une procédure (BASE) choisissant aléatoirement un nombre fixe de fins de phrase comme bornes de segment a servi de référence basse. Ses résultats dans le Tableau 1 sont des moyennes sur 1.000 tirages au sort. TOPICOLL_1 est le système décrit au paragraphe 4. TOPICOLL_2 est le même système mais avec inactivation de la partie détection de liens thématiques. Les résultats de ces deux variantes indiquent que la recherche de liens entre segments ne dégrade pas significativement les résultats de TOPICOLL en matière de segmentation. TEXTTILING est notre implé-

Systèmes	**Rappel**	**Précision**	**F1-mesure**	**Faux négatif**	**Fausse alarme**	**P_k**
SEGCOHLEX (Ferret, 1998)	0,68	0,37	0,48	nc	nc	nc
SEGAPSITH (Ferret, Grau, 2000)	0,92	0,52	0,67	nc	nc	nc
TEXTTILING	0,72	0,81	0,76	nc	nc	nc
BASE	0,51	0,28	0,36	0,46	0,55	0,50
TOPICOLL_1	0,86	0,74	0,80	0,17	0,24	0,21
TOPICOLL_2	0,86	0,78	0,81	0,17	0,22	0,20

Tableau 1 : Précision/rappel et P_k pour le corpus du *Monde*[4]

mentation de l'algorithme de Hearst en utilisant les valeurs standards de ses paramètres. Le Tableau 1 montre clairement que TOPICOLL obtient de meilleurs résultats qu'un système tel que SEGCOHLEX se fondant seulement sur un réseau de collocations. Cet avantage existe aussi par rapport aux systèmes tels que TEXTTILING qui s'appuient sur la seule récurrence lexicale et travaillent comme TOPICOLL à partir d'un contexte local. Ces éléments semblent indiquer que l'utilisation conjointe de collocations et de la récurrence lexicale est une approche intéressante. Ceci est d'ailleurs confirmé par le fait que TOPICOLL est aussi plus précis que des systèmes manipulant des représentations de thème, à l'image de SEGAPSITH. Ses performances sont également légèrement supérieures à celles obtenues par (Bigi *et al.*, 1998) en exploitant de telles représentations dans un cadre probabiliste : précision de 0,75, rappel de 0,80 et f1-mesure de 0,77 sur un corpus constitué d'articles du *Monde, a priori* différents des nôtres.

---

[3] P_k évalue la probabilité que deux mots choisis aléatoirement dans un texte 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 des segments au niveau du corpus de référence.

[4] La précision est définie par N_c / N_b et le rappel par N_c / D, où N_b est le nombre de bornes trouvées par TOPICOLL, N_c est le nombre de bornes trouvées correctes, *i.e.* correspondant à des frontières de textes dans un intervalle de 9 mots pleins autour de cette frontière, et D le nombre total de frontières de texte.

5.1.2 Évaluation de la segmentation pour l'anglais : corpus de Choi

Pour l'anglais, nous avons utilisé un corpus construit par Choi (Choi, 2000) pour comparer des systèmes de segmentation thématique. Ce corpus est composé de 700 textes artificiels constitués chacun de 10 segments, chaque segment étant formé des *n* premières phrases de documents issus du corpus *Brown*. Les sept premières lignes du Tableau 2 proviennent des

Systèmes	n ∈ [3,11]	n ∈ [3,5]	n ∈ [6,8]	n ∈ [9,11]
base - Choi	0,45	0,38	0,39	0,36
CWM (Choi, 2001)	0,09	0,10	0,07	0,05
U00 (Utiyama, Isahara, 2001)	0,10	0,09	0,07	0,05
C99 (Choi, 2000)	0,12	0,11	0,09	0,09
DOTPLOT (Reynar, 1998)	0,18	0,20	0,15	0,12
SEGMENTER (Kan *et al.*, 1998)	0,36	0,23	0,33	0,43
TEXTTILING - Choi	0,46	0,44	0,43	0,48
$TOPICOLL_1$	0,30	0,28	0,27	0,34
$TOPICOLL_2$	0,31	0,28	0,28	0,34

Tableau 2 : P_k pour le corpus de Choi

expériences réalisées par Choi (Choi, 2001). La procédure de base partitionne systématiquement chaque document en 10 segments de même longueur. Le Tableau 2 confirme que la détection de liens thématiques n'altère pas les capacités de segmentation de TOPICOLL. Il montre également que les résultats de TOPICOLL sur ce corpus sont significativement inférieurs à ceux obtenus sur le corpus du *Monde*. Une des causes possibles de cette différence réside dans notre réseau de collocations pour l'anglais : sa densité, c'est-à-dire le rapport entre la taille de son vocabulaire et son nombre de collocations, est inférieure de 30% à celle du réseau pour le français, ce qui a certainement un impact significatif. Le Tableau 2 montre aussi que TOPICOLL, qui n'utilise qu'un contexte local, a des performances inférieures à des systèmes tels que CWM, U00, C99 ou DOTPLOT, qui traitent globalement les textes qui leur sont soumis. Cette vue globale améliore la précision mais se traduit par une plus grande complexité algorithmique. Par ailleurs, la détection de liens thématiques permise par les contextes thématiques rend TOPICOLL fonctionnellement plus riche.

5.2 Évaluation globale

L'évaluation globale d'un système tel que TOPICOLL se heurte à un problème : une référence pour les liens thématiques est nécessairement liée à une segmentation de référence. Or, projeter cette référence sur les segments définis par le système à évaluer n'est pas une opération directe. Pour contourner ce problème, nous avons choisi une méthode proche de celle adoptée dans TDT pour la tâche Link Detection : nous évaluons la probabilité d'une erreur dans la classification de chaque couple de positions d'un texte comme faisant partie du même thème (Cp_{ident}) ou appartenant à des thèmes différents (Cp_{diff}). Un faux négatif est comptabilisé lorsque les positions d'un couple sont supposées relever de thèmes différents alors qu'ils sont relatifs au même thème. Une fausse alarme correspond au cas complémentaire.

Systèmes	Faux négatif	Fausse alarme	Erreur (P_k)
base	0,85	0,06	0,45
TOPICOLL	0,73	0,01	0,37

Tableau 3 : Mesures globales pour le corpus du *Monde*

Le nombre de couples Cp_{diff} étant en général beaucoup plus grand que le nombre de couples Cp_{ident}, nous avons aléatoirement sélectionné un nombre de couples Cp_{diff} égal au nombre de couples Cp_{ident} de manière à préserver une plage de valeurs assez étendue. Le Tableau 3 montre les résultats de TOPICOLL pour cette mesure et les compare à une procédure de base choisissant aléatoirement un nombre fixe de bornes de segments et de liens d'identité thématique entre segments. Cette mesure est une première proposition qui doit encore être améliorée, en particulier pour obtenir un meilleur équilibre entre les faux négatifs et les fausses alarmes.

6 Conclusion

Nous avons proposé une méthode réalisant de façon intégrée la segmentation thématique de textes et la détection de liens d'identité thématique en utilisant conjointement des collocations et la récurrence lexicale. Son évaluation a montré l'intérêt de cette approche pour des systèmes travaillant avec un contexte local. Afin d'élargir sa validation, nous envisageons de l'appliquer à des méthodes se fondant sur une appréhension globale des textes. Par ailleurs, nous souhaitons étendre son évaluation en améliorant la mesure globale que nous avons proposée et en confrontant nos résultats à des jugements humains.

Références

Beeferman D., Berger A., Lafferty J. (1999), Statistical Models for Text Segmentation, *Machine Learning*, Vol. 34(1/3), pp. 177-210.

Bigi B., de Mori R., El-Bèze M., Spriet T. (1998), Detecting topic shifts using a cache memory, Actes de la *5ème International Conference on Spoken Language Processing*, 2331-2334.

Church K. W., Hanks P. (1990), Word Association Norms, Mutual Information, And Lexicography, *Computational Linguistics*, Vol. 16(1), pp. 177-210.

Choi F., Wiemer-Hastings P., Moore J. (2001), Latent Semantic Analysis for Text Segmentation, Actes de *NAACL'01*, 109-117.

Choi F. (2000), Advances in domain independent linear text segmentation, Actes de *NAACL'00,* 26-33.

Ferret O., Grau B. (2000), A Topic Segmentation of Texts based on Semantic Domains, Actes de *ECAI 2000*, 426-430.

Ferret O. (1998) How to thematically segment texts by using lexical cohesion?, Actes de *ACL-COLING'98*, 1481-1483.

Hearst M. (1997), TextTiling: Segmenting Text into Multi-paragraph Subtopic Passages, *Computational Linguistics*, Vol. 23(1) , pp. 33-64.

Kan M-Y., Klavans J., McKeown K. (1998), Linear segmentation and segment significance, Actes du *6ème Workshop on Very Large Corpora*, 197-205.

Kaufmann S. (1999), Cohesion and Collocation: Using Context Vectors in Text Segmentation, Actes de *ACL'99*, 591-595.

Kozima H. (1993), Text Segmentation Based on Similarity between Words, Actes de *ACL'93*, 286-288.

Morris J., Hirst G. (1991), Lexical Cohesion Computed by Thesaural Relations as an Indicator of the Structure of Text, *Computational Linguistics*, Vol. 17(1) , pp. 21-48.

Passonneau R., Litman D. (1997), Discourse Segmentation by Human and Automated Means, *Computational Linguistics*, Vol. 23(1) , pp. 103-139.

Reynar R. (1998), *Topic segmentation: Algorithms and applications*, Ph.D. thesis, Computer and Information Science, University of Pennsylvania.

Utiyama M., Isahara H. (2001), A Statistical Model for Domain-Independent Text Segmentation, Actes de *ACL 2001*, 491-498.