Utilisation de relations sémantiques pour améliorer la segmentation thématique de documents télévisuels

Camille Guinaudeau, Guillaume Gravier, Pascale Sébillot

TALN 2010, Montréal, 19–23 juillet 2010

Utilisation de relations sémantiques pour améliorer la segmentation thématique de documents télévisuels*

Camille Guinaudeau[1] Guillaume Gravier[2] Pascale Sébillot[3]
INRIA Rennes[1] & IRISA (CNRS[2], INSA[3]), France
camille.guinaudeau@irisa.fr, guillaume.gravier@irisa.fr, pascale.sebillot@irisa.fr

Résumé. Les méthodes de segmentation thématique exploitant une mesure de la cohésion lexicale peuvent être appliquées telles quelles à des transcriptions automatiques de programmes télévisuels. Cependant, elles sont moins efficaces dans ce contexte, ne prenant en compte ni les particularités des émissions TV, ni celles des transcriptions. Nous étudions ici l'apport de relations sémantiques pour rendre les techniques de segmentation thématique plus robustes. Nous proposons une méthode pour exploiter ces relations dans une mesure de la cohésion lexicale et montrons qu'elles permettent d'augmenter la F1-mesure de +1.97 et +11.83 sur deux corpus composés respectivement de 40h de journaux télévisés et de 40h d'émissions de reportage. Ces améliorations démontrent que les relations sémantiques peuvent rendre les méthodes de segmentation moins sensibles aux erreurs de transcription et au manque de répétitions constaté dans certaines émissions télévisées.

Abstract. Topic segmentation methods based on a measure of the lexical cohesion can be applied as is to automatic transcripts of TV programs. However, these methods are less effective in this context as neither the specificities of TV contents, nor those of automatic transcripts are considered. The aim of this paper is to study the use of semantic relations to make segmentation techniques more robust. We propose a method to account for semantic relations in a measure of the lexical cohesion. We show that such relations increase the F1-measure by +1.97 and +11.83 for two data sets consisting of respectively 40h of news and 40h of longer reports on current affairs. These results demonstrate that semantic relations can make segmentation methods less sensitive to transcription errors or to the lack of repetitions in some television programs.

Mots-clés : Segmentation thématique, documents oraux, cohésion lexicale, relations sémantiques.

Keywords: Topic segmentation, spoken document, lexical cohesion, semantic relations.

* Travaux partiellement financés par le projet Quaero.

1 Introduction

Les travaux présentés dans cet article se placent dans le contexte de la structuration automatique de flux télévisuels et, plus particulièrement, dans ce que l'on peut considérer comme la première étape nécessaire à cette structuration : la segmentation. Afin de permettre aux utilisateurs de naviguer de façon non linéaire à l'intérieur d'un document télévisé, il est en effet essentiel de découper ce flux en émissions d'une part, et d'extraire de ces émissions des segments thématiquement cohérents d'autre part.

La segmentation de documents télévisuels quelconques ne pouvant s'appuyer sur l'utilisation des seuls indices audio disponibles – aucune méthode fondée sur ces derniers ne fournissant de résultats satisfaisants – la prise en compte d'indices alternatifs (textuels ou vidéo) apparaît nécessaire. Les performances des systèmes de reconnaissance de la parole (RAP) s'étant considérablement améliorées ces dernières années (Ostendorf *et al.*, 2008), la segmentation thématique de documents oraux peut désormais s'effectuer par le biais des transcriptions automatiques. La plupart des travaux développés en ce sens appliquent généralement sur ces transcriptions des méthodes issues de la segmentation de documents textuels, très fréquemment fondées sur la notion de cohésion lexicale. Ainsi (Mulbregt *et al.*, 1999) et (Utiyama & Isahara, 2001) proposent respectivement une technique utilisant un modèle de Markov caché et une méthode consistant à rechercher la meilleure segmentation parmi toutes les segmentations possibles. Des marqueurs discursifs, obtenus lors d'une phase préalable d'apprentissage, peuvent aussi servir à repérer des frontières thématiques ((Beeferman *et al.*, 1999) et (Christensen *et al.*, 2005)). Cependant, lors de précédents travaux sur la segmentation d'émissions radiophoniques par une approche non supervisée fondée sur la cohésion lexicale, nous avons constaté un gros écart de performances entre les segmentations de transcriptions manuelles et automatiques. En effet, les transcriptions automatiques possèdent certaines particularités. Premièrement, ces données ne contiennent ni ponctuation ni majuscule ; elles ne sont donc pas structurées en phrases comme un texte classique mais en unités appelées groupes de souffle, qui correspondent à la parole prononcée par un locuteur entre deux respirations. De plus, le taux d'erreur de notre système de RAP, même s'il reste raisonnable sur des émissions comme les journaux télévisés (JT), peut atteindre 70% pour des émissions telles que des films ou des *talk shows*, rendant impossible l'utilisation de certains indices tels que les marqueurs discursifs. Ces écarts de performances sont dûs à la qualité de l'enregistrement – enregistrement studio ou extérieur –, à la présence ou non de bruits de fond, d'applaudissements, à la différence de style de parole. Enfin, les émissions télévisuelles sont composées de segments thématiques pouvant être très courts, contenant peu de répétitions de vocabulaire (notamment au sein des journaux télévisés) et dans lesquels le niveau de langage peut être très variable (alternance présentateur/interview). Nous souhaitons donc adapter les méthodes de segmentation thématique fondées sur la cohésion lexicale à ces données particulières.

Pour pallier les difficultés liées aux erreurs de transcription, certains travaux ont proposé d'ajouter à la seule notion de cohésion lexicale des indices propres aux documents oraux. Par exemple, (Amaral & Trancoso, 2003) exploite la détection de locuteur afin de repérer le présentateur du journal télévisé, celui-ci introduisant de nouveaux reportages et donc les changements thématiques. Conjointement à la transcription, les auteurs de (Stolcke *et al.*, 1999) utilisent quant à eux la prosodie. Cependant de tels indices sont globalement peu employés car leur extraction automatique est difficile. De plus, ils permettent uniquement de remédier aux erreurs de transcription et ne traitent en rien le manque de répétitions inhérent à un corpus télévisuel. Pour rendre la segmentation thématique plus robuste aux spécificités des transcriptions automatiques d'émissions télévisées, l'utilisation de relations sémantiques nous semble pertinente, un mot mal transcrit ayant peu de chance d'être lié sémantiquement aux autres mots du segment. Certains travaux, comme (Ferret, 2002), ont intégré dans le passé des relations sémantiques dans des méthodes de segmentation de l'écrit inspirées de *TextTiling* (Hearst, 1997), c'est-à-dire basées sur la détection de

ruptures de la cohésion au sein d'une fenêtre glissante. Cependant, la méthode proposée dans (Utiyama & Isahara, 2001), fondée sur une *mesure de la cohésion* lexicale d'un segment plutôt que sur la détection de ruptures, donne de meilleurs résultats pour la segmentation de documents oraux. Nous proposons donc ici une technique originale pour introduire des relations sémantiques dans la méthode de calcul de la cohésion lexicale décrite dans (Utiyama & Isahara, 2001). Nous exploitons cette intégration dans deux méthodes de segmentation thématique, l'une globale, l'autre locale. Ces techniques, testées sur deux corpus oraux composés de documents télévisuels transcrits d'une durée globale de 80 heures (40 heures de JT et 40 heures d'émissions de reportages), permettent de découper nos données en segments thématiques composés des reportages éventuellement précédés d'un plateau de lancement. Les deux corpus utilisés constituent à notre connaissance le plus gros volume de données télévisuelles testé jusqu'à présent.

Dans cet article, nous présentons le calcul de la cohésion lexicale, d'abord sans, puis avec prise en compte des relations sémantiques, avant de décrire, en section 3, les algorithmes de segmentation que nous utilisons pour traiter nos corpus. Le choix des relations à intégrer constituant un problème majeur – elles peuvent, en effet, être sélectionnées de différentes façons, en ne conservant que les relations correspondant aux forces d'association entre mots les plus élevées par exemple ou en ne retenant qu'un nombre fixe de relations par mot – nous exposons, dans la quatrième partie, les techniques retenues d'acquisition et de sélection des relations sémantiques. Nous testons l'intégration des relations sur nos deux corpus et décrivons les résultats de ces expériences en section 5, avant la présentation de quelques perspectives.

2 Mesure de cohésion lexicale...

Le critère de cohésion lexicale fait référence aux relations lexicales qui existent au sein d'un texte et lui donnent une certaine unité. Les méthodes de segmentation, utilisant cette notion pour découper un texte en segments présentant une homogénéité du point de vue de leurs thèmes, se fondent sur l'analyse de la distribution des mots au sein du texte pour détecter des ruptures ou, de manière duale, des segments homogènes. Nous présentons, dans cette partie, une technique de mesure de la cohésion lexicale pour la segmentation de documents textuels, ainsi que la méthode que nous utilisons pour intégrer des relations sémantiques afin de rendre ce critère plus robuste aux particularités de données télévisuelles transcrites, en particulier le manque de répétitions et les erreurs de transcription.

2.1 sans prise en compte des relations sémantiques

Dans (Utiyama & Isahara, 2001), les auteurs présentent une méthode de mesure de la cohésion lexicale fondée sur le calcul d'une probabilité généralisée. La valeur de la cohésion lexicale d'un segment S_i est vue comme la mesure de la capacité d'un modèle de langue Δ_i – c'est-à-dire une distribution de probabilités – appris sur le segment S_i à prédire les mots contenus dans le segment. Cette définition de la cohésion lexicale nécessite de calculer, dans un premier temps, un modèle de langue Δ_i pour chaque segment S_i du texte à segmenter, puis de déterminer la probabilité des mots du segment S_i, étant donné Δ_i.

Modèle de langue Un modèle de langue *n-gramme* est un modèle probabiliste qui assigne une probabilité à toute séquence de n mots d'un texte. Le modèle de langue utilisé pour le calcul de la cohésion lexicale est un modèle *unigramme*, qui détermine la probabilité d'apparition de chaque mot plein – c'est-à-dire ici les noms, les adjectifs et les verbes – au sein du texte. Lors de l'estimation du modèle de langue Δ_i d'un segment S_i, on évalue la probabilité d'apparition de chacun des mots du vocabulaire du texte dans le segment S_i. Afin d'éviter que toute la masse de probabilité soit attribuée aux seuls mots apparaissant dans le segment, on applique un lissage à ce modèle de langue dans le but de redistribuer une partie des probabilités aux mots non observés – le nombre de mots observés dans le segment étant relativement petit

au regard du nombre de mots dans le texte. Le calcul du modèle de langue du segment S_i se formalise par

$$\Delta_i = \{P_i(u) = \frac{C_i(u)+1}{z_i},\ \forall\ u \in V_K\}\ \ , \tag{1}$$

avec V_K le vocabulaire, de taille K, du texte et $C_i(u)$ le compte du mot u. Dans le cas habituel, le compte d'un mot correspond à son nombre d'occurrences dans le segment S_i. La distribution de probabilités est lissée en incrémentant le compte de chacun des mots de 1. On a donc $z_i = K + \sum_{u \in V_K} Ci(u)$.

Vraisemblance La seconde étape du calcul de la cohésion lexicale d'un segment consiste à évaluer une probabilité traduisant à quel point le modèle de langue Δ_i permet d'expliquer les mots contenus dans le segment S_i, soit

$$\ln(P(S_i|\Delta_i)) = \sum_{j=1}^{n_i} \ln(\frac{C_i(w_j^i)+1}{z_i})\ \ , \tag{2}$$

avec n_i le nombre de mots dans le segment et w_j^i le j^e mot du segment. Intuitivement cette probabilité favorise les segments les plus cohérents lexicalement puisque sa valeur est plus importante lorsque les mots apparaissent plusieurs fois au sein du segment et qu'elle atteint sa valeur minimale lorsque tous les mots du segment sont différents.

Le calcul de la cohésion lexicale tel que nous venons de le présenter se base uniquement sur la répétition des mots au sein d'un texte et n'accorde aucune importance au fait que deux mots différents peuvent être sémantiquement proches. C'est pourquoi nous proposons une méthode d'intégration de relations sémantiques dans le calcul du critère de cohésion lexicale afin de le rendre moins sensible aux problèmes liés aux transcriptions automatiques.

2.2 avec prise en compte des relations sémantiques

L'intégration de relations sémantiques que nous proposons se fonde sur l'idée que si le mot « voiture », par exemple, apparaît dans un texte thématiquement homogène, alors les probabilités d'apparition des mots « conduire » ou « automobile » sont plus importantes que celles de mots n'appartenant pas au même champ lexical. Nous intégrons donc les relations sémantiques au niveau du calcul du modèle de langue de façon à ce que, pour chaque mot w_j^i rencontré dans le texte, son compte $C_i(w_j^i)$ soit incrémenté ainsi que celui des mots qui lui sont sémantiquement liés, proportionnellement à la valeur de leur proximité sémantique avec le mot w_j^i. Plus formellement, on a pour chaque mot w_j^i du segment S_i :

$$\begin{array}{l} C_i(w_j^i) = C_i(w_j^i) + 1 \\ C_i(v) = C_i(v) + r(v, w_j^i) \qquad \forall v \in V_K \qquad v \neq w_j^i\ \ , \end{array} \tag{3}$$

avec $r(v, w_j^i) \in [0, 1]$ la proximité sémantique des mots v et w_j^i, dont le calcul est décrit en section 4.

3 Segmentation thématique

Notre objectif n'est pas de mettre au point une nouvelle technique de segmentation thématique, mais de voir si des méthodes état de l'art sur du texte écrit peuvent tirer profit de l'intégration de relations sémantiques pour traiter des transcriptions automatiques. Nous utilisons donc d'une part la technique de segmentation proposée dans (Utiyama & Isahara, 2001) et, d'autre part, une technique locale dérivée de *TextTiling*, les deux s'appuyant sur une mesure de la cohésion lexicale. L'utilisation de méthodes aux comportements différents nous permet de vérifier que l'intégration de relations rend le calcul de la cohésion lexicale plus robuste aux problèmes liés aux transcriptions automatiques, indépendamment de la méthode de segmentation employée.

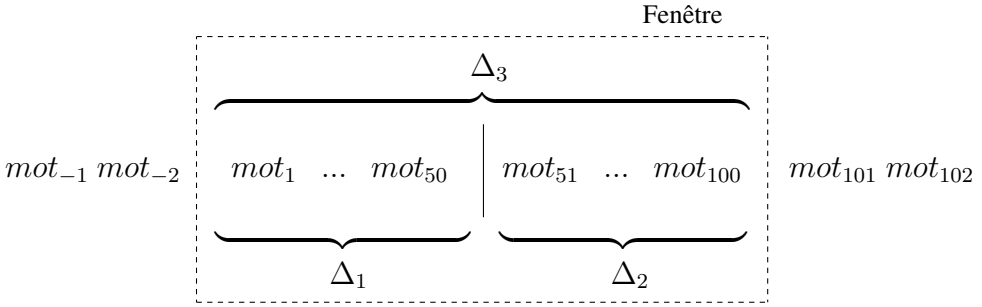

FIG. 1 – Rapport de vraisemblance généralisé

Méthode globale La méthode de segmentation développée par Utiyama et Isahara consiste à rechercher la segmentation qui produit les segments les plus cohérents d'un point de vue lexical, tout en respectant une distribution *a priori* de la longueur des segments. Son principe est de trouver la segmentation la plus probable d'une séquence de l unités élémentaires (mots, phrases, ou groupes de souffle) $W = W_1^l$ parmi toutes les segmentations possibles, soit

$$\hat{S} = \underset{S}{\operatorname{argmax}}\, P[W|S]P[S] \ . \tag{4}$$

En supposant que $P[S_1^m] = n^{-m}$, avec n le nombre de mots du texte et m le nombre de segments, la probabilité d'un texte W pour une segmentation $S = S_1^m$ est donnée par

$$\hat{S} = \underset{S_1^m}{\operatorname{argmax}} \sum_{i=1}^{m} (\ln(P[S_i|\Delta_i]) - \alpha \ln(n)) \ . \tag{5}$$

La cohésion lexicale $\ln(P[S_i|\Delta_i])$ pour le segment S_i est calculée comme décrit en section 2. Le facteur α permet de contrôler la taille moyenne des segments retournés.

Méthode locale La seconde technique de segmentation considérée est adaptée de la méthode *TextTiling* (Hearst, 1997), qui consiste à évaluer, pour chaque fenêtre centrée sur une frontière potentielle, la similarité entre la partie droite et la partie gauche de la fenêtre. Afin de comparer les deux méthodes de segmentation, globale et locale, le calcul de la similarité, fondé sur une mesure cosinus dans la technique originale, a été modifié pour utiliser la même mesure de cohésion lexicale que pour la méthode globale. Contrairement à la méthode *TextTiling*, notre méthode de segmentation locale ne consiste donc pas à comparer les parties gauche et droite de la fenêtre pour déterminer s'il y a ou non rupture de la cohésion lexicale mais plutôt à calculer le rapport de probabilité entre l'hypothèse considérant une frontière et celle n'en considérant pas (*cf.* figure 1).

Ce rapport R, s'il est important, traduit le fait que la cohésion lexicale au sein de la fenêtre est meilleure si le segment mot_1 ... mot_{100} est divisé en deux, c'est-à-dire que les vocabulaires des segments mot_1 ... mot_{50} et mot_{51} ... mot_{100} sont différents. La valeur du rapport R est donnée par :

$$R = \ln(P[mot_1 \ ... \ mot_{50}|\Delta_1]) + \ln(P[mot_{51} \ ... \ mot_{100}|\Delta_2]) - \ln(P[mot_1 \ ... \ mot_{100}|\Delta_3]) \ , \tag{6}$$

avec $\ln(P[mot_i \ ... \ mot_{i+n}|\Delta_i])$ la cohésion lexicale du segment mot_i ... mot_{i+n} calculée comme décrit en section 2.

La segmentation thématique du texte est finalement obtenue à partir des valeurs du rapport de vraisemblance pour chaque séparation entre deux groupes de souffle. De ces valeurs sont extraits des maxima locaux par une technique d'extraction des pics dominants qui, s'ils sont supérieurs à un seuil σ, définissent une frontière thématique. Cette méthode permet d'obtenir une valeur de P_k-mesure d'environ 9% sur le corpus de Choi (pour des segments composés de 9 à 11 phrases). Les résultats sont donc bien meilleurs que ceux obtenus par *TextTiling* (P_k-mesure de 48%) sur le même corpus.

4 Acquisition et sélection de relations sémantiques

Le choix des relations sémantiques intégrées dans le calcul de la cohésion lexicale peut avoir une influence importante sur les résultats de la segmentation thématique. Nous présentons ici, les méthodes d'extraction et de sélection utilisées pour obtenir les relations les plus pertinentes pour notre tâche de segmentation.

Acquisition de relations sémantiques Bien que de nombreuses ressources lexicales déjà contruites soient disponibles, elles sont malheureusement souvent liées à certain domaine. Les documents télévisuels étudiés étant multi-domaines, nous avons donc extrait les relations sémantiques à partir de corpus. L'objectif de cet article étant d'étudier l'influence de l'intégration de relations sémantiques et non d'optimiser leur extraction, nous avons choisi d'appliquer des méthodes standards afin d'acquérir deux types de relations : des relations syntagmatiques et des relations paradigmatiques. Les relations syntagmatiques correspondent à des relations de successivité et de contiguïté que les mots entretiennent au sein d'une phrase (exemple : « conduire » et « voiture »). Pour les calculer, nous avons retenu deux indices de force d'association couramment utilisés : l'information mutuelle *IM* et l'information mutuelle au cube IM^3 (Daille, 1994), ce second indice ayant été défini afin de réduire l'importance associée aux événements rares par *IM*. Le second type de relation réunit deux mots présentant une composante commune importante du point de vue du sens, comme « voiture » et « automobile ». Ces relations paradigmatiques sont calculées en associant à chaque couple de mots le cosinus de l'angle entre les vecteurs de voisinage des occurrences des deux mots. Nous obtenons ainsi une liste de synonymes, d'hypéronymes, *etc.*, non différenciés.

Pour l'ensemble des méthodes d'acquisition, les relations ont été extraites sur un corpus composé d'articles *du Monde*, *de l'Humanité* et des transcriptions de référence des campagnes *Ester 1* et *Ester 2* correspondant respectivement à 100 et 150 heures de journaux radiophoniques. Dans ce corpus, lemmatisé et normalisé, seuls les noms, adjectifs, et verbes autres que « être », « avoir » et « falloir » ont été conservés. Les scores d'association ont finalement été normalisés afin d'obtenir des valeurs comprises entre 0 et 1.

Sélection de relations sémantiques La question qui se pose alors est de retenir, pour notre tâche de segmentation, les relations sémantiques les plus pertinentes parmi tous les liens extraits. Nous explorons deux méthodes de sélection :

– conserver les δ relations ayant les scores les plus élevés tous mots confondus ($Total_\delta$) ; ces relations sont appelées « premières relations » dans la suite de cet article. Dans nos tests, la valeur de δ peut être égale à 5 000, 10 000, 20 000, 50 000 et 90 000 ;
– conserver un nombre fixe β de relations pour chaque mot ($ParMot_\beta$), β prenant les valeurs 2, 3 et 10.

De plus, nous avons remarqué que certains mots du corpus, comme « aller », « an », *etc.*, étaient sémantiquement liés avec un nombre important d'autres mots. Afin d'éviter de créer des liens sémantiques entre de trop nombreux couples de mots, ce qui conduirait à créer un nombre excessif de liens entre les segments dans nos techniques de segmentation, nous avons défini une technique de filtrage, $Seuil_\gamma$, pouvant être associée aux deux méthodes de sélection. Elle consiste à ignorer les relations sémantiques des mots qui entretiennent un nombre de relations supérieur à un certain seuil. La valeur du seuil est égale au nombre moyen de relations associées aux mots du texte à segmenter multiplié par un paramètre γ prenant les valeurs 1, 2, 3, 5 ou 10.

Le tableau 1 illustre les 5 relations aux scores d'association les plus élevés rattachées au mot « cigarette » en sélectionnant les 90 000 premières relations, $Total_{90000}$, et 10 relations par mot, $ParMot_{10}$. Nous constatons que la qualité de ces relations change selon la méthode utilisée pour leur extraction. Ainsi, les relations obtenues grâce au score *IM* correspondent généralement à des événements rares, à tel point d'ailleurs qu'aucune des relations existant avec le mot « cigarette » n'apparaît pas dans les 90 000 pre-

TAB. 1 – Relations aux scores d'association les plus élevés pour le mot « cigarette »

	IM	IM^3	paradigmatique
$Total_{90000}$		cigarette fumer cigarette paquet	cigarette cigare cigarette gitane cigarette gauloise
$ParMot_{10}$	cigarette chevignon cigarette liggett cigarette altadi cigarette détaxer	cigarette fumer cigarette paquet cigarette allumer cigarette contrebande	cigarette cigare cigarette gitane cigarette gauloise cigarette clope

mières relations, alors que les relations IM^3 et paradigmatiques semblent plus pertinentes.

5 Résultats

L'intégration des relations sémantiques pour la segmentation de documents télévisuels a été testée sur deux corpus. Le premier est constitué de 60 journaux télévisés, d'une durée de 40 minutes chacun, diffusés en février et mars 2007 sur la chaîne de télévision France 2. Le second est composé de 12 émissions de reportage de 2 heures, « Envoyé Spécial », diffusées sur France 2 entre mars 2008 et janvier 2009, et de 16 émissions de reportage « Sept à Huit », de 1 heure chacune, programmées sur la chaîne TF1 entre septembre 2008 et février 2009. Ces deux corpus ont été définis pour prendre en compte les différences importantes existant entre les deux types d'émissions : nombre de répétitions dans les JT beaucoup moins important que dans les émissions de reportages, avec une proportion de parole spontanée plus élevée dans ces derniers, et longueur moyenne des reportages d'« Envoyé Spécial » beaucoup plus grande.

Ces émissions ont été transcrites par un système de reconnaissance automatique de la parole, implémenté pour la transcription de journaux radiophoniques, atteignant un taux d'erreur d'environ 20% sur les données du corpus *Ester 2*. Les deux corpus transcrits sont composés respectivement de 12 000 et 11 000 mots pleins. Pour chacune des transcriptions, nous avons supprimé la partie précédant le lancement du premier reportage, c'est-à-dire la partie constituée des titres du journal ou du sommaire de l'émission, ainsi que celle suivant la fin du dernier reportage, ces deux parties très spécifiques perturbant l'algorithme de segmentation. Cette extraction manuelle aurait pu être effectuée en utilisant des indices audiovisuels. Une segmentation de référence a été effectuée en considérant un changement de thème à chaque changement de reportage, bien que ce ne soit pas toujours le cas, notamment pour le premier corpus. En effet, les premiers reportages des JT traitent généralement du titre principal du journal et abordent donc tous le même thème. Nous obtenons un total de 1180 frontières thématiques pour le premier corpus et de 141 pour le second. L'évaluation de nos méthodes de segmentation se fait en considérant comme correcte une frontière éloignée de moins de 10 secondes d'une frontière de référence. Nous utilisons les métriques précision, rappel et F1-mesure pour chiffrer les résultats de nos algorithmes. Afin de confronter nos différentes méthodes, nous comparons la valeur de la F1-mesure pour des valeurs de paramètre (α pour la méthode globale et σ pour la méthode locale) conduisant à une segmentation dont la longueur moyenne des segments est la plus proche de celle de la segmentation de référence.

Nous présentons dans un premier temps les résultats obtenus sur le premier corpus grâce à la méthode de segmentation globale, en analysant les techniques de sélection utilisées pour chaque type de relations. Puis, nous comparons le comportement des deux méthodes de segmentation face à l'intégration des relations sémantiques. Enfin, nous analysons la portabilité de l'intégration des relations sur le second corpus.

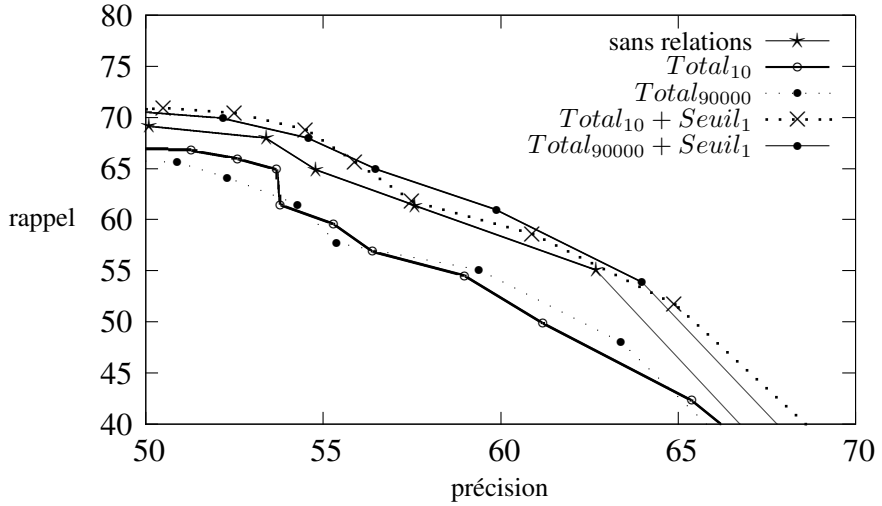

FIG. 2 – Courbes rappel/précision pour l'intégration de relations *IM*^{3} dans la méthode de segmentation globale sur des journaux télévisés, obtenues en faisant varier le facteur α contrôlant la taille des segments

5.1 Comparaison des techniques de sélection des relations sémantiques

Les deux techniques de sélection, éventuellement combinées à une méthode de filtrage, ayant été testées sur trois types de relations, nous avons obtenu un nombre important de résultats qu'il n'est pas possible de présenter ici en détails. Nous décrivons donc dans cette partie l'influence générale des techniques de sélection sur les résultats de la segmentation globale.

Information mutuelle (*IM*) L'intégration de relations sémantiques extraites par *IM* au sein de la méthode de segmentation globale permet d'améliorer les résultats lorsque la technique de sélection *ParMot* est utilisée, amélioration proportionnelle au nombre de relations introduites. La méthode de sélection *Total* ne permet pas de faire évoluer la qualité de la segmentation, de façon positive ou négative. Ceci s'explique par le fait que les toutes premières relations calculées sur nos corpus d'apprentissage par le score *IM* correspondent à des événements rares (*cf.* tableau 1) qui n'apparaissent pas dans nos transcriptions et n'ont donc pas d'influence sur la valeur de la cohésion lexicale. Finalement, l'association de la technique filtrant les mots liés à trop d'autres mots à la méthode *ParMot* permet une amélioration supplémentaire de la F1-mesure.

Information mutuelle au cube (*IM*^{3}) Lorsqu'elles sont appliquées sur des relations *IM*^{3}, les deux techniques de sélection *Total* et *ParMot* obtiennent des résultats équivalents : toutes deux détériorent à la fois les valeurs de rappel et de précision (*cf.* figure 2), détérioration d'autant plus marquée que le nombre de relations introduites est important. Cependant, l'utilisation de la méthode de filtrage $Seuil$ permet d'améliorer la qualité de la segmentation, et la meilleure valeur de la F1-mesure est obtenue grâce à la combinaison $Total_{90000} + Seuil_1$. L'introduction de trop nombreuses relations sémantiques semble donc relier un nombre excessif de mots et de segments, faisant diminuer la qualité des résultats de la segmentation et rendant nécessaire une limitation du nombre de relations introduites par filtrage.

Relations paradigmatiques L'introduction de relations paradigmatiques dans la méthode de segmentation globale a un comportement assez similaire à celui observé pour les relations *IM*^{3}. En effet, les résultats de la segmentation se dégradent avec le nombre croissant de relations introduites et ceci, pour les deux techniques *Total* et *ParMot*. À nouveau, le filtrage des mots auxquels on associe des relations sémantiques permet d'améliorer les résultats. Enfin, les résultats pour les deux techniques de sélection sont, pour ces relations également, tout à fait équivalents, même si la meilleure valeur de la F1-mesure est ici obtenue

TAB. 2 – Valeurs[1] de la F1-mesure pour l'intégration des relations sémantiques

	Méthode globale		Méthode locale	
Corpus	JT	reportages	JT	reportages
Sans relations	59.44	51.09	26.19	26.62
IM	$ParMot_{10} + Seuil_5$ 61.41	$ParMot_{10} + Seuil_5$ 61.42	$ParMot_{10}$ 26.29	$ParMot_{10} + Seuil_1$ 39.20
IM^3	$Total_{90000} + Seuil_1$ 60.44	$Total_{5000} + Seuil_1$ 62.92	$Total_{90000}$ 27.11	$ParMot_2 + Seuil_2$ 39.27
Paradigmatiques	$ParMot_{10} + Seuil_3$ 61.27	$ParMot_3 + Seuil_3$ 62.28	$ParMot_{10} + Seuil_3$ 28.02	$Total_{90000} + Seuil_2$ 41.54

avec une méthode de sélection différente, $ParMot_{10} + Seuil_3$.

5.2 Comparaison des méthodes de segmentation globale et locale

Bien qu'une augmentation ponctuelle de la valeur de la F1-mesure est obtenue pour l'intégration des trois types de relations dans la méthode de segmentation locale, l'ajout de relations sémantiques n'a ici que peu d'influence sur la qualité des segmentations obtenues, comme nous avons pu l'observer sur les courbes rappel/précision. Deux facteurs peuvent expliquer ces résultats. Tout d'abord, lors de l'ajout de relations sémantiques au sein de la fenêtre glissante, si les valeurs de la cohésion lexicale des trois segments – ceux situés à droite et à gauche de la fenêtre et celui constitué de tous les mots de la fenêtre – sont augmentées, l'intégration des relations n'aura que très peu d'effet sur la valeur du rapport R calculé. Par ailleurs, la technique d'extraction des maxima locaux utilisée pour définir les frontières thématiques peut également permettre d'interpréter ces résultats. En effet, cette technique, très sensible aux variations, peut proposer des segmentations différentes alors même que l'évolution de la valeur de la cohésion lexicale au fil du texte présente des profils fortement similaires. Enfin, nous pouvons constater dans le tableau 2 que les résultats de la méthode locale sont bien moins élevés que ceux de la méthode globale. Ceci peut s'expliquer par le fait que notre corpus de JT est constitué de segments de tailles très variables et pouvant être très petits. La fenêtre glissante peut alors contenir plusieurs segments, ce qui rend la valeur de la cohésion lexicale au sein de cette fenêtre inexploitable, la variabilité de la taille des segments au sein d'un même corpus empêchant de définir une taille de fenêtre optimale.

5.3 Portabilité de l'intégration de relations sémantiques sur d'autres corpus

En analysant les courbes rappel/précision pour l'ajout de relations sémantiques dans la méthode locale, nous constatons que les résultats fournis sur le corpus de reportages sont peu encourageants bien que, ponctuellement, la valeur de la F1-mesure soit augmentée pour les trois types de relations (tableau 2). Concernant la méthode globale, l'intégration d'un grand nombre de relations paradigmatiques et *IM*^{3} dégrade, ici aussi, la qualité de la segmentation sauf si on associe un filtrage aux méthodes de sélection. De plus, le comportement des relations *IM* est, lui aussi, similaire puisque la sélection *Total* ne permet pas d'influencer la qualité de la segmentation et qu'une sélection des relations par mot est nécessaire pour obtenir un impact sur les résultats. Cependant, l'amélioration obtenue lors de l'intégration des relations sémantiques sur ce corpus est plus sensible que celle observée sur les JT et ceci pour les deux méthodes. Cette différence s'explique par le fait que le nombre de répétitions est ici plus important. À chaque répé-

[1]Mesures obtenues pour des valeurs de paramètres conduisant à une segmentation pour laquelle la longueur des segments est la plus proche de celle de la segmentation de référence.

tition d'un mot, des relations sémantiques vont être intégrées, renforçant la valeur de la cohésion lexicale pour les segments cohérents. L'augmentation de la valeur de la F1-mesure doit toutefois être analysée avec prudence car le nombre de frontières dans ce corpus est beaucoup plus faible.

6 Conclusion

Nous avons proposé une méthode pour intégrer des relations sémantiques dans une mesure de la cohésion lexicale et montré que l'utilisation de ces relations dans une méthode de segmentation thématique globale améliore les résultats pour des documents télévisuels transcrits, compensant en partie l'absence de répétitions dans les documents considérés et les erreurs de transcription. Nous avons mis en évidence l'importance de limiter le nombre de relations introduites afin d'empêcher qu'elles ne relient entre eux un nombre excessif de segments. Afin d'améliorer nos méthodes de segmentation de documents oraux, nous privilégions deux pistes utilisant des caractéristiques des transcriptions automatiques. La première exploite les mesures de confiance associées à chacun des mots de la transcription – correspondant à la probabilité que le mot soit correctement transcrit – afin de mieux gérer les erreurs de transcription. L'introduction de ces seules mesures dans de précédents travaux, nous a permis d'améliorer la F1-mesure de +1.5 sur le corpus de journaux télévisés. En associant relations sémantiques et mesures de confiance, nous espérons combiner leurs gains et rendre ainsi notre méthode plus adaptée aux transcriptions automatiques. La seconde piste consiste à fonder notre segmentation non pas sur une seule transcription, c'est-à-dire la meilleure hypothèse fournie par le système de reconnaissance automatique de la parole, mais sur la liste des n meilleures hypothèses qu'il propose.

Références

AMARAL R. & TRANCOSO I. (2003). Topic indexing of TV broadcast news programs. In *6th International Workshop on Computational Processing of the Portuguese Language*.

BEEFERMAN D., BERGER A. & LAFFERTY J. (1999). Statistical models for text segmentation. *Machine Learning*, **34**(1-3), 177–210.

CHRISTENSEN H., KOLLURU B. & ET *al.* Y. G. (2005). Maximum entropy segmentation of broadcast news. In *30th IEEE ICASSP*.

DAILLE B. (1994). *Approche mixte pour l'extraction automatique de terminologie : statistiques lexicales et filtres linguistiques*. PhD thesis, Université de Paris 7.

FERRET O. (2002). Segmenter et structurer thématiquement des textes par l'utilisation conjointe de collocations et de la récurrence lexicale. In *9e Actes de Traitement Automatique des Langues Naturelles*.

HEARST M. A. (1997). Texttiling : segmenting text into multi-paragraph subtopic passages. *Computational Linguistics*, **23**, 33–64.

MULBREGT P. V., CARP I., GILLICK L., LOWE S. & YAMRON J. (1999). Segmentation of automatically transcribed broadcast news text. In *DARPA Broadcast News Workshop*.

OSTENDORF M., FAVRE B. & ET *al.* R. G. (2008). Speech segmentation and spoken document processing. *IEEE Signal Processing Magazine*, **25**(3), 59–69.

STOLCKE A., SHRIBERG E. & ET *al.* D. H.-T. (1999). Combining words and speech prosody for automatic topic segmentation. In *DARPA Broadcast News Workshop*.

UTIYAMA M. & ISAHARA H. (2001). A statistical model for domain-independent text segmentation. In *9th ACL*.