Segmentation thématique : processus itératif de pondération intra-contenu

Abdessalam Bouchekif[1,2], Géraldine Damnati[1], Delphine Charlet[1]
(1) Orange Labs , 2, Avenue Pierre Marzin 22307 Lannion Cedex
(2) Laboratoire d'Informatique de l'Universite du Maine, LIUM - France
{abdessalam.bouchekif,geraldine.damnati,delphine.charlet}@orange.com

RÉSUMÉ

Dans cet article, nous nous intéressons à la segmentation thématique d'émissions télévisées exploitant la cohésion lexicale. Le but est d'étudier une approche générique, reposant uniquement sur la transcription automatique sans aucune information externe ni aucune information structurelle sur le contenu traité. L'étude porte plus particulièrement sur le mécanisme de pondération des mots utilisés lors du calcul de la cohésion lexicale. Les poids TF-IDF sont estimés à partir du contenu lui-même, qui est considéré comme une collection de documents mono-thème. Nous proposons une approche itérative, intégrée à un algorithme de segmentation, visant à raffiner la partition du contenu en documents pour l'estimation de la pondération. La segmentation obtenue à une itération donnée fournit un ensemble de documents à partir desquels les poids TF-IDF sont ré-estimés pour la prochaine itération. Des expériences menées sur un corpus couvrant différents formats des journaux télévisés issus de 8 chaînes françaises montrent une amélioration du processus global de segmentation.

ABSTRACT

An iterative topic segmentation algorithm with intra-content term weighting

This paper deals with topic segmentation of TV Broadcasts using lexical cohesion. The aim is to propose a generic approach, only relying on the automatic speech transcription with no external nor a priori information on the TV content. The study focuses on a new weighting scheme for lexical cohesion computation. TF-IDF weights are estimated from the content itself which is considered as a collection of mono-thematic documents. We propose an iterative process, integrated to a segmentation algorithm, aiming to refine the partition of a content into documents in order to estimate the weights. Topic segmentation obtained at a given iteration provides a set of documents from which TF-IDF weights are re-estimated for the next iteration. An experiment on a rich corpus covering various formats of Broadcast News shows from 8 French TV channels improves the overall topic segmentation process.

MOTS-CLÉS : Segmentation thématique, pondération TF-IDF, cohésion lexicale, TextTiling
KEYWORDS : Topic segmentation, TF-IDF weighting, lexical cohesion, TextTiling

1 Introduction

La segmentation thématique consiste à effectuer un pavage d'un document (texte classique, audio ou vidéo) en segments thématiquement homogènes. Plusieurs programmes de recherche se sont attachés à traiter la segmentation thématique de journaux télévisés (JT) mais le problème demeure d'actualité et doit être considéré avec de nouvelles perspectives

pour pouvoir traiter des contenus à la ligne éditoriale de plus en plus variée. En particulier, la structuration traditionnelle d'un JT où le présentateur principal, en plateau, introduit un nouveau sujet suivi d'un reportage ou d'une interview, tend à être substituée ou complétée par des mises en scènes plus modernes. Dans certains JT sont intercalées des brèves lues par le présentateur principal ou par un autre journaliste, sans qu'un reportage ne vienne illustrer le propos (c'est le cas du journal d'Arte par exemple). Au contraire, certains JT contiennent une succession de reportages, sans retours plateaux et sans introduction par le présentateur principal (c'est le cas du journal du soir de France 3 qui inclut en fin de programme une succession de reportages issus des éditions régionales, ainsi que certains JT de M6 ou d'Euronews qui n'ont pas du tout de présentateur principal). La plupart des études dans la littérature ont porté sur des corpus de JT de format traditionnel. Une des particularités du présent travail est d'être mené sur un corpus varié de JT issus de 8 chaînes différentes, de durée et de format divers.

Dans la littérature, trois catégories d'indices ont été exploitées : des indices lexicaux, acoustiques et visuels. La combinaison de ces indices est en règle générale profitable à la tâche de segmentation (Wang et al., 2012). Cependant, les deux derniers sont fortement liés aux règles éditoriales de chaque chaîne télévisée (Xie et al. 2010) : présence ou non d'un présentateur principal, présence ou non de titres incrustés ou de logos.

Notre objectif étant de développer un système de segmentation thématique générique, nous avons fait le choix de privilégier les indices lexicaux qui révèlent des frontières à partir de variations sémantiques dans un contenu, indépendamment de toute sorte d'information structurelle sur l'émission traitée. L'exploitation spécifique d'informations structurelles peut améliorer les performances comme dans (Bouchekif et al., 2013), mais nous cherchons ici à améliorer en amont l'approche générique basée sur la cohésion lexicale.
Plusieurs algorithmes de segmentation thématique basés sur la cohésion lexicale ont été proposés dans la littérature (voir par exemple (Eisenstein et Barzilay, 2008) pour une revue des approches). Les algorithmes varient tant du point de vue de la méthode de détection des frontières que du point de vue de la mesure de similarité (y compris des approches en recrudescence à base de Latent Semantic Analysis). Même si l'approche de TEXTTILING (Hearst, 1997) initialement conçue pour segmenter du texte s'est avérée peu performante sur des contenus audiovisuels (Claveau et Lefèvre, 2011) (Guinaudeau et al., 2010), nous avons néanmoins choisi d'adopter ce schéma de façon à en explorer deux dimensions. La première est la méthode de sélection des frontières à partir de la courbe de similarité et la seconde est le mécanisme de pondération des mots utilisé pour calculer une valeur pertinente de cohésion lexicale. Ce choix n'est néanmoins pas restrictif et les propositions développées dans cet article peuvent s'appliquer à des algorithmes plus sophistiqués.
L'article est structuré de la façon suivante : la section 2 présente notre algorithme de segmentation thématique, la section 3 présente une évolution vers une approche intégrée itérative qui permet de raffiner la pondération TF-IDF des mots. Les expériences sont présentées dans la section 4.

2 Algorithme de segmentation thématique

Comme pour l'algorithme TEXTTILING, la similarité est calculée entre chaque paire de blocs adjacents. Les segments unitaires considérés sont des groupes de souffle (GS), c'est-à-dire des séquences de mots entre deux pauses dans un tour de parole. Les pauses et les changements de locuteur sont détectés automatiquement par le système de transcription

automatique. La similarité est donc calculée tout au long de l'émission à l'aide d'une fenêtre glissante de taille K, entre des blocs adjacents de K GS de part et d'autre de la frontière potentielle. Il en résulte une courbe de cohésion lexicale à partir de laquelle sont extraites les hypothèses de frontières. Dans les deux premières sous-sections, nous décrivons comment sont réalisés la pondération des termes et le calcul de similarité lexicale. Nous proposons ensuite un algorithme de sélection des frontières à partir de la courbe de cohésion obtenue.

2.1 La pondération TF-IDF intra-document

La pondération TF-IDF est largement utilisée en recherche d'information (RI) pour évaluer le pouvoir discriminant d'un terme t dans un document d (via TF : fréquence locale du terme), relativement à une collection de documents (via IDF : fréquence globale inverse du terme). Dans le cadre de la segmentation thématique, la pondération des mots permet d'augmenter la pertinence des mesures de similarité lexicale, en renforçant la contribution de certains mots dans l'estimation de ces mesures. Dans le domaine de la segmentation de contenus du type information (journaux télévisés, journaux radiophoniques, émissions de reportages), les poids sont généralement estimés par un large corpus. Par exemple, (Guinaudeau et Hirschberg, 2011) utilisent l'outil *kiwi* (Lecorvé et al., 2008) qui produit des poids estimés à partir d'une collection de 800000 articles du journal *Le Monde*. Afin de nous affranchir de la contrainte de disposer d'une base d'apprentissage, nous nous proposons de suivre l'approche donnée dans (Malioutov et al., 06) où les auteurs introduisent une pondération intra-document pour le domaine de la segmentation thématique de conférences. Sans aucune information externe, les poids TF-IDF sont estimés uniquement à partir du contenu en question. Le principe est de découper uniformément l'émission en N morceaux (ou *chunk*). Chaque *chunk* est une succession de groupes de souffles et correspond à l'équivalent d'un document en RI. Le terme t dans le groupe de souffle x est associé au poids $w(c(x), t)$ qui dépend du *chunk* $c(x)$ dans lequel se trouve x.

$$w(c(x), t) = TF_{c(x),t} \times IDF_t, \text{ où } IDF_t = \log\left(\frac{N}{n_t}\right) \qquad (1)$$

où $TF_{c(x),t}$ est la fréquence du terme t dans le morceau $c(x)$ et n_t est le nombre de *chunks* dans lequel le terme t apparaît.

Cette approche permet de faire ressortir les mots discriminants dans un passage de l'émission relativement aux autres passages. Des expériences utilisant d'autres pondérations comme Okapi n'ont pas permis d'améliorer les performances de la segmentation.

2.2 Calcul de similarité

La mesure cosinus permet de mesurer la proximité entre la représentation vectorielle de deux blocs adjacents b_j et b_{j+1}. Le coefficient associé au terme t dans la représentation vectorielle d'un bloc b est une valeur pondérée $v(b, t)$. Dans notre approche, il n'y a pas unicité de la pondération TF-IDF dans un bloc donné car les groupes de souffles du bloc peuvent ne pas appartenir tous au même *chunk*. Ainsi le coefficient associé à t dans le bloc b est obtenu en sommant les fréquences pondérées du terme t dans chaque GS du bloc :

$$v(b, t) = \sum_{x \in b} \left(f_{x,t} \times w(c(x), t) \right) \qquad (2)$$

où $f_{x,t}$ est la fréquence du terme t dans le GS x.
Pour une frontière potentielle j entre deux blocs b_j et b_{j+1}, la similarité est donnée par

$$cohesion(j) = \frac{\sum_t \left(v(b_j,t) \times v(b_{j+1},t)\right)}{\sqrt{\sum_t \left(v(b_j,t)\right)^2} \times \sqrt{\sum_t \left(v(b_{j+1},t)\right)^2}}. \quad (3)$$

Le nombre de chunks *N* est calculé automatiquement pour chaque émission en fonction de sa durée et de la durée moyenne des thèmes de l'ensemble d'émissions.

2.3 Algorithme de division récursive (Splitting)

Plusieurs stratégies ont été introduites pour sélectionner les frontières à partir de la courbe de similarité. L'approche classique (Hearst, 1997) consiste à détecter les vallées (un point entouré par deux pics) et à calculer leur profondeur en faisant la somme des deux différences (entre le point et le pic à gauche d'une part et le point et le pic à droite d'autre part). Les vallées dont la profondeur dépasse un certain seuil (approche dite par *seuillage*) sont considérées comme des points de transition thématique. Il faut noter que les points qui ne correspondent pas à des vallées valent 0. Il peut se produire que plusieurs vallées profondes apparaissent dans un court intervalle de temps, ou bien qu'un changement thématique se traduise par une succession de vallées de profondeur limitée. (Lu et al., 2011) proposent une approche basée sur la programmation dynamique pour optimiser globalement la recherche des frontières dans la courbe. (Claveau et Lefèvre, 2011) ont proposé d'appliquer en plus d'une métrique alternative basée sur la vectorisation, l'algorithme dit Ligne de Partage des Eaux (LPE) issu de la morphologie mathématique pour réaliser un partitionnement de l'émission à partir de la courbe.

Pour améliorer la robustesse de l'extraction des frontières, nous proposons un nouvel algorithme avec deux particularités : premièrement nous proposons d'exploiter conjointement la similarité lexicale et la profondeur des vallées, et deuxièmement nous avons implémenté un algorithme itératif de partitionnement d'une émission à partir de la courbe. Il résulte de la première observation que la recherche directe sur les valeurs de similarité n'est pas optimale (certains changements de thèmes entre deux sujets proches, sur un même pays par exemple, peuvent se traduire par une similarité relativement importante). De façon similaire, travailler uniquement sur la profondeur des vallées n'est pas optimal : (les pics de part et d'autre peuvent ne pas être très hauts si un sujet ne contient que peu de répétitions de termes). Nous proposons ainsi de combiner ces deux mesures complémentaires à l'aide d'une interpolation linéaire. Pour une frontière potentielle *j*, le score suivant doit être maximisé :

$$score(j) = \lambda \left(1 - cohesion(j)\right) + (1 - \lambda)\, depth\,(j)) . \quad (4)$$

La deuxième proposition est un algorithme de division récursive lors duquel est définie une zone d'exclusion autour des frontières trouvées à chaque itération. Le partitionnement consiste à construire un ensemble S de segments :

1. Initialement, S contient un seul segment constitué de l'émission entière.
2. Chaque segment de S est coupé en deux, le point de coupure correspond à la valeur maximale du score, si cette valeur dépasse un seuil donné.
3. Les GS situés autour du point de coupure ne seront pas pris en considération lors de la prochaine itération.
4. Les segments obtenus sont présentés à l'étape 2.

L'étape 3 permet de limiter les phénomènes de maxima locaux et garantit que l'on n'obtiendra pas plusieurs frontières consécutives. La zone de neutralisation est fixée à 3 GS

de part et d'autres d'une frontière. L'algorithme s'arrête lorsqu'aucun point de coupure candidat ne dépasse le seuil. Cette approche par zone d'exclusion s'est avérée plus efficace qu'un lissage de la courbe pour limiter l'effet des maxima locaux. La granularité des groupes de souffles est trop grande pour envisager un lissage efficace sans perte d'information.

3 Pondération itérative intra-document

Dans cette section, nous introduisons une variation de la pondération TF-IDF intra-document. Le principe initial présenté dans la section 2.1 consistait à découper uniformément le contenu en *N chunks* simulant la notion de document. Au-delà de ce découpage uniforme, nous proposons une approche itérative, utilisant les résultats de notre algorithme de segmentation thématique pour déterminer les chunks. La segmentation obtenue à une itération donnée fournit un ensemble de documents à partir desquels les poids TF-IDF sont ré-estimés pour l'itération suivante.

Initialement, le document est coupé en *N* morceaux uniformes. Le nombre de chunks N est obtenu automatiquement pour chaque émission en divisant la durée de l'émission par une durée moyenne de segments thématiques estimée sur un corpus de développement. L'indice du premier GS de chaque *chunk* uniforme est considéré comme l'ensemble initial de frontières et est placé dans le vecteur hyp_0. A l'itération i, les hypothèses de l'itération $i-1$ (hyp_{i-1}) sont utilisées pour estimer les pondérations TF-IDF (i). La combinaison linéaire entre la cohésion lexicale et la profondeur des vallées est recalculée. Ensuite, l'algorithme de division récursive est appliqué pour déterminer les hypothèses hyp_i .

L'algorithme s'arrête lorsque la segmentation se stabilise (pas de changement significatif entre les hypothèses de deux itérations successives hyp_i et hyp_{i+1}). Afin de mesurer objectivement cette stabilisation, nous utilisons la mesure p_k (Beeferman et al., 1999).
La mesure p_k compare une segmentation de référence R et une hypothèse de segmentation *H*. Elle est basée sur le principe d'une fenêtre glissante de taille *k* parcourant la totalité de l'émission. Dans cette métrique, la tâche de segmentation est vue comme un problème de classification binaire répondant à la question : le groupe de souffle *j* et le groupe de souffle *j+k* appartiennent-ils au même segment ? p_k mesure la probabilité que deux GS distants de k soient classés de la même façon par *R* et par *H*.

$$p_k(R,H) = \frac{1}{n-k}\sum_{j=1}^{n-k} f\left(f(r_j, r_{j+k}), (h_j, h_{j+k})\right) \qquad (5)$$

La fonction f vaut 1 si ses deux arguments sont identiques, sinon elle vaut 0. Plusieurs études ont mis en exergue qu'une faible valeur de k favorise la précision de cette mesure. Dans notre implémentation, la valeur de k est fixée à 6. L'algorithme s'arrête lorsque la valeur de p_k entre hyp_{i-1} et hyp_i est proche de 1 $\left(1 - p_k(hyp_{i-1},\ hyp_i)\right) \leq \epsilon$.

Il faut noter que nous n'avons pas encore étudié la preuve de la convergence de l'algorithme. L'algorithme s'arrête au bout de 6 itérations si le critère d'arrêt n'a pas atteint la valeur du seuil ε. En pratique trois à quatre itérations sont suffisantes.

4 Expériences et résultats

Les expériences sont menées sur deux corpus d'émissions. Le premier pour le développement (Dev) est constitué de 33 JT de 7 chaînes françaises (TF1, France2, France3,

LCI, France24, Arte, M6). Le deuxième pour le test (*Test*) est composé de 6 JT d'une autre chaîne : Euronews. La particularité de ce corpus est qu'il n'y a ni présentateur principal ni plateau, il s'agit uniquement d'une succession de reportages, chacun associé à un reporter différent et de longueur variable. Le tableau 1 résume les caractéristiques de ces deux corpus.

	Dev	*Test*
Nombre d'émissions	33	6
Durée moyenne	$\sim 22\ min$	$\sim 26\ min$
Nombre de frontières (par JT)	397 (11,5)	156 (26,0)
Durée moyenne des thèmes	$115\ s$	$79\ s$

TABLE 1 – Description des corpus

Ces émissions ont été transcrites à l'aide du moteur de reconnaissance automatique de la parole de Vocapia Research basé sur le système du LIMSI (Gauvain et al., 2002). Le taux d'erreurs mots sur le corpus de *Dev* est de 16,1%. Nous ne pouvons donner les performances du corpus *Test* qui n'a pas été transcrit manuellement. Les mots qui ont un score de confiance inférieur à 0.5 sont écartés. Les pré-traitements classiques ont été appliqués : lemmatisation (Lia_tagg : http://pageperso.lif.univ-mrs.fr/~frederic.bechet/download.html), suppression de certains mots non porteurs de sens à partir d'une stop-liste. Par ailleurs, de façon similaire à (Guinaudeau et al., 2010) nous avons écarté la première partie d'un journal lorsqu'elle ne contient que des titres et la dernière partie lorsqu'il s'agit du rappel des titres. La définition d'un segment est plus large que la simple notion de reportage, il s'agit d'un segment thématiquement cohérent. L'annotation manuelle a été validée par deux annotateurs. Pour les cas ambigus (comme un long passage dans un journal relatant diverses conséquences des chutes de neiges ou dela crise économique), nous avons fait le choix de segmenter en sujet pour chaque conséquence.

La taille de la fenêtre pour la détermination de la courbe de cohésion lexicale a été optimisée sur le Dev et a été fixée à 16 groupes de souffles. Le coefficient λ d'interpolation pour le calcul du score a été fixé à 0,75. Le seuil ε sur la mesure p_k pour le critère d'arrêt de l'algorithme itératif vaut 0,09. Les performances sont mesurées en termes de rappel et de précision, en comparant la segmentation de référence avec celle d'hypothèse. De façon similaire à plusieurs travaux dans la littérature, une tolérance de 10 *s* a été autorisée entre les frontières d'hypothèses et de références.

La figure 1 illustre l'importance de l'algorithme de sélection et du score de similarité. Les courbes rappel/précision ont été obtenues en faisant varier le seuil présent dans chacune des approches. Les 6 courbes de la figure 1 représentent la combinaison des trois scores (cohésion, profondeur de vallée et l'interpolation des deux) et des deux algorithmes de sélection (*seuillage* et *splitting*). La pondération TF-IDF est calculée selon une partition uniforme de l'émission en N chunks. *Seuillage* ($vallee$) correspond à l'algorithme de TEXTTILING de base. *Splitting* ($cohesion + vallee$) correspond à notre proposition. La combinaison linéaire des deux scores associée à l'algorithme de *splitting* augmente la F-max de 12,5 points. On observe plus de fausses alarmes (faible précision) avec l'approche *seuillage* (toutes les hypothèses dépassant le seuil sont prises en compte), contrairement à l'approche *splitting* avec une zone d'exclusion qui permet d'avoir une meilleure précision dans les zones de plus fort rappel. Les trois courbes correspondant à l'approche *splitting* montrent clairement que l'interpolation de la cohésion et la profondeur des vallées est une bonne stratégie, avec une F-max finale de 55,3% sur le Dev.

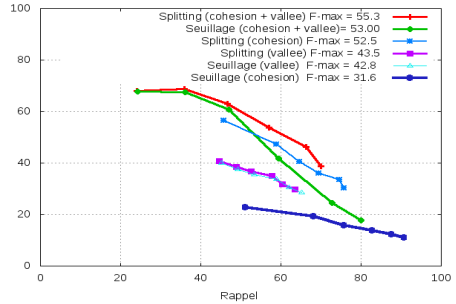

FIGURE 1 – Impact de la stratégie de sélection et de calcul du score

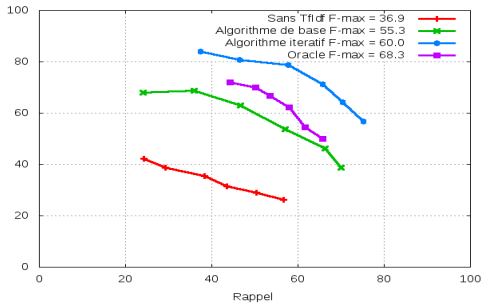

FIGURE 2 – Impact de la stratégie itérative de pondération.

Pour le reste des expériences, *Splitting* (*cohesion* + *vallee*) est systématiquement utilisé. Lorsque nous utilisons la pondération TF-IDF avec le découpage uniforme pour les chunks, cette technique sera dénommée *baseline*. Afin de montrer l'intérêt de la pondération itérative, nous comparons notre approche à une version dans laquelle aucune pondération ne serait utilisée (similarité calculée uniquement à partir de la fréquence) et à une version où la pondération TF-IDF serait calculée à partir de la segmentation thématique réelle, obtenue à partir des frontières de référence (condition Oracle). Les résultats de la figure 2 montrent que les meilleures performances sont obtenues dans les conditions Oracle, avec une importante marge de progression par rapport à la baseline (de 55,3% à 68,3% de F-max) confortant ainsi le potentiel de notre proposition. Les performances sont sérieusement dégradées lorsqu'aucune pondération n'est appliquée.

Le système itératif améliore la *baseline*, permettant de passer d'une F-max de 55,3% à une F-max de 60,0%. Près de 30% de l'écart entre la condition *baseline* et la condition Oracle a pu être comblé. Enfin, nous avons validé l'approche sur un nouvel ensemble de JTs issus d'Euronews. Le tableau 2 illustre les résultats obtenus sur ce corpus en choisissant le seuil optimal établi de façon à atteindre la F-max sur le corpus de Dev.

Condition de pondération	*Rappel*	*Précision*	$F-mesure$
Algorithme de base	46,8	59,3	52,3
Itératif	53, 8	62,7	57,9
Oracle	57,7	69,8	63,2

TABLE 2 – Résultats sur le Test (Euronews)

Les mêmes tendances peuvent être observées, avec une meilleure couverture de l'écart entre la *baseline* et l'Oracle. En analysant les résultats, il s'avère que l'algorithme itératif permet de retrouver des frontières entre des sujets proches (deux sujets sportifs, deux reportages consécutifs sur un même pays). Une évaluation plus fine de la pondération TF-IDF permet de remonter ce type de frontières difficilement accessibles pour les approches lexicales.

5 Conclusion

Cet article propose une approche de segmentation thématique s'appuyant uniquement sur le contenu lexical d'un Journal Télévisé. L'objet de notre approche est de développer une méthode générique pouvant s'appliquer à tout type de JT, indépendamment de sa structure.

A partir d'un algorithme classique de segmentation thématique basé sur la cohésion lexicale locale (TEXTTILING), deux modifications sont proposées : la première portant sur le processus d'extraction des frontières thématiques à partir de la courbe de cohésion et la seconde portant sur l'estimation des poids associés aux mots pour l'estimation de cette cohésion. Les deux propositions se traduisent par une amélioration significative des performances de segmentation sur un corpus diversifié de JT issus de 8 chaînes. L'approche itérative de calcul de la pondération TF-IDF à partir du contenu lui-même n'est pas limitée à notre algorithme mais peut avoir une portée beaucoup plus large dans de nombreux contextes d'utilisation.

Références

BEEFERMAN, D., BERGER, A., et LAFFERTY, J. D. (1999). Statistical models for text segmentation, *Machine Learning*, pages 177–210.

BOUCHEKIF, A., DAMNATI, G., et CHARLET, D. (2013). Complementarity of Lexical Cohesion and Speaker Role Information for Story Segmentation of French TV Broadcast News. *In Proc. of SLSP.*

CLAVEAU, V., et LEFEVRE, S. (2011). Segmentation thématique : apport de la vectorisation, *Actes de la conférence CORIA*.

EISENSTEIN, J. et BARZILAY, R. (2008). Bayesian Unsupervised Topic Segmentation, *In Proc. EMNLP*.

GAUVAIN, J.L., LAMEL. et ADDA, G. (2002) The LIMSI Broadcast News Transcription System. *Speech Communication*, pages 89-108.

GUINAUDEAU C., GRAVIER G. et SÉBILLOT P. (2010). Utilisation de relations sémantiques pour améliorer la segmentation thématique de documents télévisuels, *In Proc. TALN*

GUINAUDEAU, C., et HIRSCHBERG, J. (2011). Accounting for prosodic information to improve asr-based topic tracking for TV Broadcast. *In Proc. of Interspeech*.

HEARST, M. (1997). TextTiling: segmenting text into multiparagraph subtopic passages, *Computational Linguistics, pages 33–64*.

LECORVE, G., et GRAVIER, G. (2008). An unsupervised web-based topic language model adaptation method''. *In Proc. of ICASSP*.

LU, M., LEUNG, C., XIE, L., MA, B., et LI, H. (2011). Probabilistic Latent Semantic Analysis for Broadcast News Story Segmentation, *In Proc. of Interspeech*.

MALIOUTOV, I., et BARZILAY, R. (2006). Minimum cut model for spoken lecture segmentation. *In Proc. ACL,* pages 25–32.

WANG, X., XIE, L., MA, B., CHNG, E.-S. et Li. H. (2012). Broadcast News Story Segmentation Using CRF and Multi-modal Features. *IEICE Transactions on Information and Systems*, pages 1206-1215.

XIE, L., YANG, Y., LIU, Z-Q, FRENG. W. et LIUM, Z. (2010). Integrating Acoustic and Lexical Features In Topic Segmentation of Chinese Broadcast News Using Maximum Entropy Approach. *In Proc. of ICALIP*.