Indices utiles à la cohésion lexicale pour la segmentation thématique de documents oraux

Camille Guinaudeau, Guillaume Gravier, Pascale Sébillot

Indices utiles à la cohésion lexicale pour la segmentation thématique de documents oraux*

Camille Guinaudeau - Guillaume Gravier - Pascale Sébillot

IRISA UMR 6074 & INRIA Rennes - Bretagne Atlantique
Campus de Beaulieu, F - 35042 Rennes Cedex, France
camille.guinaudeau@irisa.fr – guillaume.gravier@irisa.fr – pascale.sebillot@irisa.fr

ABSTRACT

The increasing quantity of TV material requires methods to help users navigate such data streams. Topic segmentation of TV broadcast is a first stage to structuring tasks. The goal of this article is to determine to what extent confidence measures and semantics can compensate errors in automatic transcripts for topic segmentation. To this end, we introduce confidence measure and semantic relations in a topic segmentation method. We show that our F1-measure is improved by +1.5 and +1.9 when integrating confidence measure and semantic relations respectively. Such improvement demonstrates that simple clues can conteract errors in automatic transcripts for topic segmentation.

Keywords: confidence measure, semantic relations, topic segmentation, TV broadcast

1. Introduction

L'augmentation du nombre de documents télévisuels disponibles rend indispensable la mise en place de méthodes de structuration de ces flux, structuration nécessitant une phase préalable de segmentation : du flux en émissions successives d'une part et des émissions en segments thématiques d'autre part. La segmentation thématique de documents oraux peut désormais s'effectuer par le biais des transcriptions automatiques de la bande sonore contenue dans les documents, les performances des systèmes de reconnaissance automatique de la parole (RAP) s'étant considérablement améliorées ces dernières années [5]. 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 le critère de cohésion lexicale. Ainsi [4] et [7] proposent respectivement une méthode modélisant le problème de la segmentation par 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 [2, 3]. Christensen et *al.* [3] ont d'ailleurs établi que les erreurs de transcriptions n'avaient que peu d'effets sur les performances d'un algorithme de segmentation supervisée utilisant de tel marqueurs. Cependant, nous avons obtenu, lors de précédents travaux sur la segmentation d'émissions radiophoniques par une approche non supervisée se basant sur la cohésion lexicale, un gros écart de performances entre transcriptions manuelles et automatiques. Ceci laisse penser qu'il convient, pour pallier les erreurs de transcription, d'intégrer à la cohésion lexicale des indices propres aux documents oraux. Par exemple, [1] 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 [6] utilisent quant à eux la prosodie. Cependant de tel indices sont globalement peu employés car leur extraction automatique est difficile. L'objectif de cet article est donc d'étudier si des indices plus facilement accessibles peuvent aider notre algorithme de segmentation à être plus robuste aux erreurs de transcription. Pour cela, nous testons si l'intégration des mesures de confiance, associées à chacun des mots transcrits par le système de RAP, et de relations sémantiques permet d'améliorer les performances de notre méthode de segmentation.

Dans cet article, nous présentons tout d'abord le critère de cohésion lexicale tel qu'il est utilisé pour la segmentation thématique de documents textuels, avant de décrire, en section 3, les modifications apportées à ce critère pour la segmentation de documents oraux. Les méthodes de segmentation employées ainsi que leurs résultats sont exposés dans les sections 4 et 5, avant la présentation de quelques perspectives.

2. Cohésion lexicale pour l'écrit

La notion 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 thématique utilisant cette cohésion se basent sur l'analyse de la distribution des mots au sein du texte : une rupture thématique est détectée lorsque le vocabulaire utilisé change de façon significative.

La valeur de la cohésion lexicale d'un segment S_i peut être vue comme la mesure de la capacité d'un modèle de langue Δ_i appris sur le segment S_i à prédire les mots contenus dans le segment. Deux étapes importantes sont nécessaires dans le calcul de la valeur de la cohésion lexicale d'un segment :

- le calcul du modèle de langue Δ_i du segment S_i,
- le calcul de la probabilité traduisant la capacité du modèle de langue Δ_i à prédire les mots de S_i.

Modèle de langue. Le modèle de langue Δ_i appris sur un segment S_i est un modèle de langue, sur l'en-

*Travaux partiellement financés par le projet Quaero.

semble des mots du vocabulaire du texte, pour ce segment. Le calcul du modèle de langue du segment S_i se formalise, pour un lissage donné, par

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

avec V_K le vocabulaire du texte de taille K et $C_i(u)$ le compte du mot u qui correspond à son nombre d'occurrences dans le segment S_i. La distribution de probabilité est lissée en incrémentant le compte de chacun des mots de 1. L'objectif de ce lissage est d'empêcher la concentration de la masse de probabilité sur les mots observés dans le segment – le nombre de mots observés dans le segment étant relativement petit au regard du nombre de mots dans le texte. On a donc $z_i = K + \sum_{j=1}^{n_i} Ci(w_j^i)$, avec n_i le nombre de mots du segment S_i et w_j^i le j^e mot de S_i.

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

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

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 diminue si beaucoup de mots sont différents.

Le calcul de la cohésion lexicale tel que nous venons de le présenter accorde autant d'importance à chacun des mots, qu'ils soient ou non correctement transcrits. Or une erreur de transcription peut avoir un impact important sur la valeur de la cohésion lexicale.

3. Cohésion lexicale pour l'oral

Afin d'adapter le calcul de la cohésion lexicale aux documents oraux, en étant robuste aux erreurs de transcription, nous souhaitons y intégrer deux indices : des mesures de confiance et des relations sémantiques.

3.1. Utilisation des mesures de confiance

Les mesures de confiance, qui correspondent à la probabilité pour un mot d'être correctement transcrit, peuvent être prises en compte à deux niveaux lors du calcul de la cohésion lexicale : au moment du calcul du modèle de langue Δ_i ou lors du calcul de la probabilité des mots du segment S_i pour ce modèle de langue.

Dans le premier cas, on se propose de remplacer le compte $C_i(u)$ par la somme des confiances associées à chacune de ses occurrences, soit

$$C'_i(u) = \sum_{w_j^i = u} c(w_j^i)^{\lambda_2} \qquad (3)$$

où $c(w_j^i)$ correspond à la valeur de l'indice de confiance du j^e mot dans le segment S_i et où λ_2 est un paramètre permettant de faire diminuer le poids des mots dont la valeur de la mesure de confiance est faible.

Dans le second cas, la probabilité d'apparition de l'occurrence d'un mot dans un segment est multipliée par la mesure de confiance de l'occurrence de ce mot, ceci dans le but de diminuer l'importance du mot dans le calcul de la cohésion lexicale si sa mesure de confiance est faible

$$\ln(P(S_i|\Delta_i)) = \sum_{j=1}^{n_i} (c(w_j^i))^{\lambda_1} \ln(\frac{C_i(w_j^i)+1}{z_i})\ , \qquad (4)$$

avec λ_1 équivalent à λ_2.

Il est également possible de combiner les deux techniques d'intégration en remplaçant dans (4) C_i par C'_i.

3.2. Utilisation de la sémantique

Contrairement à un mot correctement transcrit, un mot mal transcrit a peu de chance d'être relié sémantiquement aux autres mots du segment. À partir de cette hypothèse, des relations sémantiques s'apparentant à des synonymes – nous considérons que deux mots sont sémantiquement liés s'ils apparaissent dans des contextes similaires – sont intégrées dans le calcul de la cohésion lexicale afin de diminuer le poids des mots mal transcrits dans ce calcul. Le compte C_i dans (1) est modifié de la manière suivante

$$C''_i(u) = C_i(u) + \beta \sum_{j=1, w_j^i \neq u}^{n_i} r(w_j^i, u)\ , \qquad (5)$$

avec $r(w_j^i, u)$ la proximité sémantique des mots w_j^i et u. Cette proximité correspond à la valeur de similarité des contextes de leurs occurrences.

Le paramètre β est utilisé pour pondérer l'importance des relations sémantiques lors du calcul du modèle de langue. Cette pondération est rendue nécessaire par le fait que les valeurs des relations sémantiques intégrées sont trop faibles par rapport au lissage et aux nombres d'occurrences.

4. Segmentation thématique

Nous présentons maintenant la méthode de segmentation thématique, basée sur le calcul de la cohésion lexicale, que nous utilisons, et la façon dont nous exploitons l'intégration des indices cités pour la segmentation de documents oraux.

4.1. Approche générale

Notre méthode de segmentation se base sur l'une des meilleures techniques de segmentation thématique existante, développée par Utiyama et Isahara [7], qui 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. Le principe de cette technique est de trouver la segmentation la plus probable d'une séquence de l unités élémentaires (mots, phrases, *etc.*) $W = W_1^l$ parmi toutes les segmentations possibles, soit

$$\hat{S} = \underset{S}{\operatorname{argmax}}\, P[W|S]P[S]\ . \qquad (6)$$

En supposant que $P[S_1^m] = n^{-m}$, la probabilité d'un texte W pour une segmentation $S = S_1^m$ est donnée par

$$\hat{S} = \underset{S_1^m}{\operatorname{argmax}} \sum_{i=1}^{m} (\ln(P[S_i|\Delta_i]) - \alpha \ln(n))\ , \qquad (7)$$

avec n le nombre de mots du texte. 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.

4.2. Approches proposées

Intégration des mesures de confiance. Afin d'adapter la méthode de Utiyama et Isahara à des documents oraux, nous définissons les 4 méthodes d'intégration suivantes :

- V^{λ_1} : l'intégration des mesures de confiance se fait lors du calcul de la vraisemblance (Éq. 4),
- M^{λ_2} : l'intégration des mesures de confiance se fait lors du calcul du modèle de langue (Éq. 3),
- $V^{\lambda_1} + M^{\lambda_2}$: l'intégration se fait lors du calcul de la vraisemblance et lors du calcul du modèle de langue,
- *Seuil* : seuls les mots dont la mesure de confiance est supérieure à un certain seuil sont pris en compte lors du calcul du modèle de langue (Éq. 1).

Cette dernière méthode est comparable à la méthode M^{λ_2}. En effet, plus la valeur du λ_2 augmente, plus la valeur de $c_i(w_j^i)^{\lambda_2}$ devient petite si $c_i(w_j^i)$ est faible. Ainsi les seuls mots pris en compte sont ceux dont la valeur de $c_i(w_j^i)$ est supérieure à un certain seuil.

Intégration de relations sémantiques. Les relations sémantiques intégrées, en modifiant le calcul du modèle de langue comme décrit dans (5), ont été apprises sur un corpus composé de textes contenant des articles *du Monde*, *de l'Humanité*, des transcriptions manuelles des campagnes *Ester 1* et *Ester 2*. Elles sont calculées en associant à chaque mot un vecteur composé des mots qui apparaissent dans ses voisinages, pondérés par leur fréquence d'apparition. La proximité sémantique de deux mots correspond à la valeur du cosinus de leurs vecteurs de contexte.

5. Résultats

Nos méthodes de segmentation prenant en compte les mesures de confiance ou les relations sémantiques ont été testées sur un corpus composé de 57 journaux télévisés (d'environ 1/2 heure chacun) diffusés en février et mars 2007 sur la chaîne de télévision France 2. Ces émissions ont été transcrites par un système de RAP, implémenté pour la transcription de journaux radiophoniques, atteignant un taux d'erreur d'environ 20% sur les journaux français du corpus *Ester 2*. Pour chacune des transcriptions, nous avons supprimé la partie précédant le lancement du premier reportage (titres) ainsi que celle suivant la fin du dernier, ces deux parties très spécifiques perturbant l'algorithme de segmentation. Cette extraction manuelle aurait pu être effectuée en utilisant des indices vidéo ou audio tels que la détection du bandeau de titre par exemple. 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 : ainsi, les premiers reportages traitent généralement du principal titre du journal et abordent donc tous le même thème. Nous obtenons ainsi un total de 1 180 frontières thématiques. 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 leurs résultats pour une valeur de α optimale, c'est-à-dire conduisant à une segmentation dont la longueur moyenne des segments est la plus proche de celle de la segmentation de référence (96.1 secondes).

Tab. 1: F1-mesure pour la méthode *Seuil*

seuil	0	0.1	0.3	0.5	0.7	0.9
F1-mesure	58.9	58.9	58.7	59.1	57.9	55.0

5.1. Intégration de mesures de confiance

Nous présentons tout d'abord la méthode *Seuil*, la plus facile et la plus immédiate à mettre en œuvre, avant de décrire les résultats des méthodes nécessitant une modification de la technique de calcul du critère de cohésion lexicale.

Dans le tableau 1, nous pouvons constater que la méthode *Seuil* ne permet pas d'améliorer significativement les performances de l'algorithme de segmentation. En effet, un seuil égal à 0.5 conduit à une très faible amélioration de la F1-mesure (+0.2) par rapport à une segmentation sans prise en compte de la mesure de confiance (seuil = 0). De plus, nous remarquons que pour les seuils dont la valeur est supérieure à 0.7, la valeur de la F1-mesure diminue fortement car le nombre de mots pris en compte lors du calcul du modèle de langue est trop faible.

Dans le tableau 2 sont résumés les résultats obtenus en intégrant les mesures de confiance grâce aux méthodes modifiant la technique de calcul de la cohésion lexicale ; la première ligne correspond à une segmentation sans prise en compte des mesures de confiance. Nous pouvons constater que ces trois méthodes ont un comportement similaire. En effet, pour toutes ces techniques, l'intégration des mesures de confiance permet d'améliorer de façon statistiquement significative (t-test) la valeur de la F1-mesure lorsque le paramètre λ_k[1] est égal à 1 ou 2, mais dégrade la qualité de la segmentation lorsque que la valeur du λ_k est trop importante. Nous remarquons également que l'amélioration est plus importante lorsque les deux méthodes sont combinées, +1.5 contre +1.2 et +1 pour les méthodes V_{λ_1} et M_{λ_2} respectivement. Finalement, nous observons que la méthode V_{λ_1} entraîne une dégradation plus importante que M_{λ_2} lorsque le paramètre λ_k augmente – constatation valable également lorsque les deux techniques sont combinées. Ceci peut s'expliquer par le fait que la probabilité d'apparition d'un mot u dans le segment S_i correspond à une probabilité lissée, ce qui implique que, même si la valeur de $c_i(w_j^i)^{\lambda_2}$ est très petite, la probabilité d'apparition du mot dans M_{λ_2} n'est pas autant diminuée que dans V_{λ_1}.

En étudiant les courbes rappel/précision pour α variant de 0 à 1, pour chacune des 4 méthodes – courbes calculées pour une valeur de λ_k optimale – nous avons pu constater que, pour un rappel supérieur à 55%, l'intégration de mesures de confiance permettait d'améliorer à la fois le rappel et la précision pour toutes les mé-

[1] avec $k = 1$ ou $k = 2$.

Tab. 2: F1-mesure pour V^{λ_1}, M^{λ_2} et $V^{\lambda_1} + M^{\lambda_2}$.

λ_k	V^{λ_1}	M^{λ_2}	$V^{\lambda_1} + M^{\lambda_2}$ avec $\lambda_1 = 1$	$V^{\lambda_1} + M^{\lambda_2}$ avec $\lambda_1 = \lambda_2$
0	59.8	59.8	59.8	59.8
1	**61.0**	60.1	**61.3**	**61.3**
2	60.1	**60.8**	60.3	60.0
3	58.3	58.5	60.2	59.2
4	57.7	59.4	60.1	58.3
5	57.1	59.7	59.8	57.7
6	56.0	59.4	59.1	56.2
7	38.7	59.1	58.6	55.5
8	38.2	59.0	58.5	54.8
9	37.9	58.3	58.6	54.1
10	37.7	57.9	58.1	53.1

thodes. De plus, nous avons remarqué que l'intégration de mesures de confiance conduit d'une part à l'augmentation du nombre de frontières correctes détectées par notre algorithme mais également au déplacement de frontières préalablement reconnues, les rapprochant ainsi de frontières de référence.

L'intégration des mesures de confiance pourrait également être effectuée en combinant la méthode *Seuil* et celles modifiant la technique de calcul de la cohésion lexicale. Cependant, un premier test sur l'association de *Seuil* et M_{λ_2} n'a pas fourni de meilleurs résultats que M_{λ_2} seule (+0.2 contre +1). Nous avons donc choisi de ne pas tester plus avant les combinaisons.

De toutes ces constations, nous pouvons conclure que les méthodes modifiant le calcul de la cohésion lexicale pour intégrer les mesures de confiance offrent une plus forte amélioration de la qualité de la segmentation thématique que la méthode *Seuil*. De plus, cette amélioration est plus importante lorsque l'intégration des mesures de confiance est effectuée lors des deux étapes nécessaires au calcul de la cohésion lexicale.

5.2. Intégration de relations sémantiques

Lors de l'intégration de relations sémantiques dans le calcul de la cohésion lexicale, nous avons tout d'abord effectué différents tests en faisant varier le paramètre β afin de donner plus ou moins de poids aux relations sémantiques. Nous avons également modifié le nombre de relations sémantiques utilisées lors du calcul de la cohésion lexicale. En effet, nous avons pu constater qu'un nombre trop important bruitait énormément ce calcul. Nous avons donc choisi de limiter le nombre de relations associées à chaque mot. Avec un paramètre β optimal, la valeur de la F1-mesure est augmentée de +1.9 lorsque le nombre de relations associées à chaque mot est égal à 2, et de +0.7 lorsqu'il est égal à 3.

De plus, en observant en détails les résultats, nous remarquons que l'intégration de relations sémantiques permet à la fois de supprimer des frontières incorrectes et, comme pour les mesures de confiance, de rapprocher certaines frontières de frontières de référence.

L'intégration de relations sémantiques permet donc, elle aussi, d'améliorer les performances de l'algorithme de segmentation. Cependant, il est nécessaire de contraindre le nombre de relations intégrées et le poids qui leur est associé, un poids trop important ou un trop grand nombre de relations faisant diminuer considérablement la valeur du rappel.

6. Conclusion et perspectives

L'intégration d'indices supplémentaires que sont les mesures de confiance et les relations sémantiques semble rendre notre algorithme de segmentation plus robuste aux erreurs de transcription et améliore ses performances. En effet, la valeur de la F1-mesure est augmentée de +1.5 dans le cas de l'intégration des mesures de confiance et de +1.9 lors de l'utilisation de relations sémantiques lorsque nous utilisons la méthode de segmentation proposée dans [7]. De plus, des résultats obtenus sur une méthode basée sur le principe de fenêtres glissantes fournit des résultats similaires, ce qui semble indiquer que l'intégration des deux indices permet d'améliorer la qualité de la segmentation quelle que soit la méthode utilisée.

Afin de consolider ces résultats, nous souhaitons tester l'intégration des deux indices sur un autre corpus composé d'émissions télévisées différentes. En effet, même si un premier test sur 4 émissions de reportages donne des résultats encourageants – la F1-mesure est augmentée de +12 lorsque l'on utilise la méthode $V^{\lambda_1} + M^{\lambda_2}$ par rapport à une segmentation sans prise en compte des mesures de confiance – le corpus n'est pas suffisamment conséquent pour nous permettre actuellement d'en tirer des conclusions. De plus, dans le but d'améliorer les performances de notre algorithme de segmentation, nous souhaitons également combiner l'intégration des mesures de confiance et des relations sémantiques lors du calcul de la cohésion lexicale. Enfin, une perspective à plus long terme consiste à appliquer notre méthode de segmentation non pas sur la transcription finale mais sur les graphes de mots.

Références

[1] R. Amaral and I. Trancoso. Topic indexing of TV broadcast news programs. In *6th International Workshop on Computational Processing of the Portuguese Language*, 2003.

[2] D. Beeferman, A. Berger, and J. Lafferty. Statistical models for text segmentation. In *Machine Learning*, 1999.

[3] H. Christensen, B. Kolluru, and Y. Gotoh et *al.* Maximum entropy segmentation of broadcast news. In *30th IEEE ICASSP*, 2005.

[4] P. Van Mulbregt, I. Carp, L. Gillick, S. Lowe, and J. Yamron. Segmentation of automatically transcribed broadcast news text. In *DARPA Broadcast News Workshop*, 1999.

[5] M. Ostendorf, B. Favre, and R. Grishman et *al.* Speech segmentation and spoken document processing. *IEEE Signal Processing Magazine*, 2008.

[6] A. Stolcke, E. Shriberg, and D. Hakkani-Tür et *al.* Combining words and speech prosody for automatic topic segmentation. In *DARPA Broadcast News Workshop*, 1999.

[7] M. Utiyama and H. Isahara. A statistical model for domain-independent text segmentation. In *9th ACL*, 2001.