

Segmentation et Titrage Automatique de Journaux Télévisés

Abdessalam Bouhekif^{1,2} Géraldine Damnati¹ Nathalie Camelin² Delphine Charlet¹ Yannick Estève²

⁽¹⁾Orange Labs, 2 avenue Pierre Marzin 22300, Lannion, France.

⁽²⁾Laboratoire d'Informatique de l'Université du Maine, LIUM - France.

Résumé. Dans cet article, nous nous intéressons au titrage automatique des segments issus de la segmentation thématique de journaux télévisés. Nous proposons d'associer un segment à un article de presse écrite collecté le jour même de la diffusion du journal. La tâche consiste à apparier un segment à un article de presse à l'aide d'une mesure de similarité. Cette approche soulève plusieurs problèmes, comme la sélection des articles candidats, une bonne représentation du segment et des articles, le choix d'une mesure de similarité robuste aux imprécisions de la segmentation. Des expériences sont menées sur un corpus varié de journaux télévisés français collectés pendant une semaine, conjointement avec des articles aspirés à partir de la page d'accueil de Google Actualités. Nous introduisons une métrique d'évaluation reflétant la qualité de la segmentation, du titrage ainsi que la qualité conjointe de la segmentation et du titrage. L'approche donne de bonnes performances et se révèle robuste à la segmentation thématique.

Abstract.

Automatic Topic Segmentation and Title Assignment in TV Broadcast News

This paper addresses the task of assigning a title to topic segments automatically extracted from TV Broadcast News video recordings. We propose to associate to a topic segment the title of a newspaper article collected on the web at the same date. The task implies pairing newspaper articles and topic segments by maximising a given similarity measure. This approach raises several issues, such as the selection of candidate newspaper articles, the vectorial representation of both the segment and the articles, the choice of a suitable similarity measure, and the robustness to automatic segmentation errors. Experiments were made on various French TV Broadcast News shows recorded during one week, in conjunction with text articles collected through the Google News homepage at the same period. We introduce a full evaluation framework allowing to measure the quality of topic segment retrieval, topic title assignment and also joint retrieval and titling. The approach yields good titling performance and reveals to be robust to automatic segmentation.

Mots-clés : Segmentation thématique, Titrage automatique, Pondération Okapi, Mesures de similarité.

Keywords: Topic segmentation, Title assignment, Okapi weighting, Similarity measures.

1 Introduction

Avec les dernières innovations technologiques, les sites des chaînes TV proposent gratuitement à leurs téléspectateurs des services de TV de rattrapage (*Replay/catch up TV*) via Internet. Ce service donne à l'utilisateur la possibilité de voir les émissions des chaînes TV à travers les podcasts. Ces derniers intéressent aussi les fournisseurs d'actualités comme Google Actualités, Yahoo Actualités, Orange Actualités, *etc.* Ils collectent des informations de différentes sources (site web des journaux, radio, chaînes TV) en agrégeant les documents de thèmes similaires. Elle permet à l'utilisateur d'avoir une information présentée par différents média. Pour permettre cette exploitation, le document doit être préalablement segmenté en fragments thématiquement homogènes : *i.e.* parlant d'un seul sujet. Cette tâche est nommée *segmentation thématique*. Les segments retournés sont identifiés par des labels anonymes (*sujet₁, sujet₂, ...*).

Donner un titre à chaque thème abordé durant l'émission est un atout supplémentaire pour une meilleure diffusion des contenus. En effet, le titre est le meilleur moyen pour décrire les différentes parties de l'émission de façon rapide et précise. D'ailleurs, ce sont les premiers éléments qu'un utilisateur consulte en priorité. Cependant, un titre doit remplir certains critères : il doit être *correct, court, clair et compréhensible*, offrant un aperçu à la fois global et spécifique du contenu. Ces critères rendent la tâche de titrage automatique plus complexe pour la machine. Le contenu même de l'émission peut contenir des informations utiles à la tâche de titrage :

- *Visuel* : un changement de sujet peut être accompagné par un titre vidéo. L'*OCR* (Optical Character Recognition) permet de détecter et de reconnaître les textes d'une vidéo quand ils sont incrustés.

- *Acoustique* : certains événements sonores, comme les jingles, peuvent aider à la détection de l'énoncé des titres par le présentateur.

- *Linguistique* : les mots les plus pertinents peuvent jouer le rôle d'un titre. Ainsi, une partie du discours contenant le plus grand nombre de mots discriminants peut être aussi considérée comme un titre.

L'extraction des titres à partir du document même n'est pas toujours possible. En effet, les indicateurs visuels et acoustiques sont fortement liés aux choix éditoriaux des chaînes télévisées (le titrage incrusté n'est pas tout le temps disponible, le présentateur principal ne donne pas forcément au début du journal tous les sujets abordés, *etc.*). Le contenu linguistique de l'émission donne des titres moins informatifs (les mots clés peuvent fournir un titre ambigu et demandent un effort supplémentaire de la part de l'utilisateur). De plus, l'extraction d'une partie du discours est un problème compliqué qui dépend du système de reconnaissance automatique de la parole et de la segmentation du discours.

Dans la littérature, malgré l'importance du titrage, peu de travaux traitent des journaux télévisés. Dans (Hsueh & Moore, 2006), les auteurs cherchent à catégoriser des réunions pour cela ils définissent un ensemble fini de catégories. La tâche de titrage est alors vue comme un problème de classification multi-classes. Les auteurs de (Lau *et al.*, 2011) proposent une méthode dans laquelle les titres des articles *Wikipédia* sont des candidats potentiels. Chaque thème est représenté par une liste de *Top10* mots, laquelle est considérée comme une requête. Ainsi, le titre du document le plus proche de la requête est considéré comme un identifiant du thème.

Notre objectif est de développer un système de titrage générique indépendant de toute information structurelle *a priori* sur l'émission tout en respectant la définition d'un bon titre. Nous proposons une nouvelle approche de titrage des segments de journaux télévisés obtenus automatiquement, qui exploite des informations provenant de la presse écrite. Elle consiste à apparier un segment à un article de presse du même jour tout en décrivant le même thème, afin d'attribuer le titre de l'article à ce segment. Cette approche permet de donner à nos segments des titres rédigés par des journalistes professionnels qui d'un point de vue thématique sont proches des segments automatiques.

L'article est structuré comme suit : dans la section 2, on décrit brièvement notre algorithme de segmentation thématique. La section 3 est consacrée à notre approche du titrage. Enfin, les expériences et la conclusion sont présentées respectivement dans les sections 4 et 5.

2 Segmentation thématique

Plusieurs algorithmes de segmentation thématique (ST) ont été proposés dans la littérature. Parmi ces algorithmes, citons *textTiling* (Hearst, 1997), *C99* (Choi, 2000) et *MinCut* (Malioutov & Barzilay, 2006). Un état de l'art plus complet est donné dans (Boucekif *et al.*, 2014b) avec également plus de détails quant à notre approche de ST. Cette approche est inspirée du *TextTiling* mais exploite plus largement les propriétés des données orales traitées. Elle repose sur l'analyse de la distribution des mots entre deux fenêtres textuelles pour déterminer l'existence ou non d'une borne thématique entre deux fenêtres adjacentes. Nous considérons comme segments unitaires les groupes de souffle (GS) qui sont les paroles prononcées par un locuteur entre deux respirations (pauses silencieuses). Ces groupes sont rassemblés en blocs de K GS pour calculer la similarité entre chaque paire de blocs adjacents. À l'aide d'une fenêtre glissante, la similarité est donc calculée tout au long de l'émission entre des blocs adjacents de K GS de part et d'autre des frontières potentielles. Il en résulte ainsi une courbe de cohésion à partir de laquelle sont extraites des frontières.

(Boucekif *et al.*, 2014a) proposent deux façons de pondérer les mots de l'émission dont le principe général est de découper l'émission en N morceaux (ou chunks) de taille différente, où chaque chunk représente la notion d'un document en recherche d'information. Dans la première méthode, le début de chaque intervention du présentateur principal représente le début d'un nouveau chunk. Dans la deuxième méthode, les poids sont calculés de manière itérative. La segmentation obtenue pour une itération donnée fournit un ensemble de documents à partir desquels les poids seront ré-estimés dans la prochaine itération.

Par ailleurs, nous avons introduit dans (Boucekif *et al.*, 2014b) la notion de la cohésion de la parole regroupant la cohésion lexicale et locuteurs dans une seule notion. Une frontière potentielle est valide si la distribution conjointe des mots et des locuteurs diffère suffisamment de part et d'autre de la frontière.

3 Titrage automatique

Notre approche de titrage intervient après la phase de ST de journaux télévisés, elle consiste à associer à chaque segment obtenu le titre de l'article de presse du jour traitant du même sujet. Parmi tous les articles de presse disponibles, le titre associé au segment est celui de l'article le plus proche thématiquement. Le titrage, tel que présenté, peut donc être vu comme une tâche d'ordonnement d'une liste de titres candidats, incluant les rejets (aucun candidat ne sera retenu).

Étape 1 : Aspirer des articles de presse

La première étape consiste à récolter tous les articles de presse parus sur la page d'accueil de Google Actualités le même jour que le segment à titrer. Les pages étant bruitées par des éléments non-informatifs, nous utilisons l'outil Boilerpipe¹ pour ne garder que le contenu utile de l'article. Toutes les informations concernant les articles sont sauvegardées dans une base de données. Chaque tuple (un article de presse) contient l'identifiant de l'article, le lien de la page web, le titre de l'article, sa date de publication, le contenu de la page et l'identifiant de l'article principal².

Étape 2 : Représentation vectorielle

Le calcul de la similarité entre segment et article de presse nécessite une représentation vectorielle de l'un et l'autre. Or le segment est issu de la transcription de la parole (contenant potentiellement des disfluences et des erreurs de reconnaissance) et l'article est composé de texte écrit dans un style journalistique. La représentation choisie doit être robuste à diverses sources et styles. Avec l'aide du logiciel *Lia-tag*, des prétraitements standards (lemmatisation, filtrage des mots) ont été appliqués sur les articles de presse et les segments. Ensuite, chaque article de presse et chaque segment est remplacé par une liste de mots pertinents selon la mesure *Okapi*. Enfin, une normalisation est réalisée relativement aux mots ayant le score le plus élevé, et un filtrage est appliqué : les mots ayant un score supérieur à +0.25 sont conservés.

Étape 3 : Calcul de similarité

Le but de cette étape est de calculer la similarité entre chaque couple (article, segment). Le choix de la mesure appropriée à la nature de nos documents est très important. Pour cela, nous proposons de comparer les mesures mentionnées dans le tableau 1 telles qu'elles sont définies dans (Curran & Moens, 2002). La mesure *Jaccard* est tout simplement le rapport entre le nombre de termes en commun et le nombre de tous les termes apparaissant dans l'article et le segment. La mesure *Lin* est une version pondérée de la mesure *Jaccard*.

Set_JACCARD	$\frac{ S \cap A }{ S \cup A }$	Extended_JACCARD	$\frac{\sum_{t \in S \cap A} w^S(t) \times w^A(t)}{\ W^S\ _2^2 + \ W^A\ _2^2 - \sum_{t \in S \cap A} w^S(t) \times w^A(t)}$
LIN	$\frac{\sum_{t \in S \cap A} w^S(t) + w^A(t)}{\sum_{t \in S \cup A} w^S(t) + w^A(t)}$	Cosine	$\frac{\sum_{t \in S \cap A} w^S(t) \times w^A(t)}{\ W^S\ _2 \times \ W^A\ _2}$

TABLE 1 – Mesures de similarité utilisées où S est la liste des termes (après filtrage) du segment, A est la liste des termes (après filtrage) de l'article considéré. On note par w_t^S (resp. w_t^A) le poids du terme t dans le segment S (resp. A)

Les valeurs de similarité sont exploitées non seulement pour classer les articles mais aussi pour filtrer les résultats. En effet, un sujet abordé dans un journal télévisé n'est pas forcément traité dans la presse. Par conséquent, le premier document retourné prend une valeur de similarité faible. L'application d'un seuil α permettra de régler ce genre de problème et de prendre uniquement les documents dont on est sûr qu'ils développent la même thématique.

4 Expériences

4.1 Corpus et annotation

Les expériences sont menées sur un corpus constitué de 86 journaux télévisés collectés durant la période du 10 au 16 février 2014 en provenance de 8 chaînes françaises (TF1, France2, france3, M6, Arte, D8, NT1, Euronews). Les émissions ont été transcrites par le biais du système Vocapia (Gauvain *et al.*, 2002) qui obtient 16,1% de taux d'erreurs (WER) sur un corpus équivalent. Dans la même période et avec un rythme d'une fois par heure, nous avons collecté à travers le site web Google Actualités une base de données de 22000 articles. Nous conservons uniquement les articles principaux³ et supprimons les doublons⁴ réduisant le volume à 4600 articles. Pour simplifier la tâche d'annotation nous exploitons seulement les articles principaux, ce qui donne en moyenne 660 articles par jour.

Le grand nombre d'articles de la collection et de segments de JT rend la tâche d'annotation manuelle très longue et fastidieuse. En effet, il faut que l'annotateur évalue la potentielle association thématique entre chaque segment et les articles de la collection. Afin de réduire le nombre d'association à évaluer, nous proposons à l'annotateur pour chaque segment, uniquement l'ensemble des articles de presse du même jour ayant au moins 2 mots en commun avec le segment considéré. Ainsi, l'annotateur a vérifié en moyenne 127 titres par segment (et non les 660 pour chaque segment).

1. <https://code.google.com/p/boilerpipe/>

2. Google Actualités regroupe les articles qui traitent du même sujet, chaque thème est représenté par un article considéré comme le principal.

3. Pour chaque article principal, plusieurs articles portant sur le même sujet sont proposés par Google.

4. Le même article peut être téléchargé plusieurs fois dans la journée.

La tâche d'annotation consiste à indiquer pour chaque article proposé : (i) si le titre de l'article reflète bien le contenu du segment, (ii) s'il résume partiellement le segment (ne couvre pas la totalité du segment, ne suit pas l'actualité, etc.), ou (iii) si le titre n'a pas de relation thématique avec le segment. Dans ce travail, uniquement les titres résumant parfaitement le contenu du segment sont considérés comme corrects et les autres comme des titres incorrects. Chaque segment de référence R est associé à un ensemble de titres candidats. Cet ensemble peut être vide si le thème n'a pas été traité dans les articles de la collection, le segment est alors considéré comme un segment *non titrable* (\bar{T}). Dans le cas contraire, il est considéré comme *titrable* (T). Au final, le corpus se compose de 658 segments de type T et 339 segments de type \bar{T} . Le

	Nb	Dur. Moy	T	\bar{T}
Seg. longs	761	131,4	467	294
Seg. courts	236	20,4	191	45
Tous les Seg.	997	105,1	658	339

TABLE 2 – Répartition des segments titrables et non titrables par rapport à la durée des segments (courts et longs)

Le tableau 2 représente la répartition des segments titrables et non titrables par rapport à la durée des segments. Deux types de segments sont considérés : *long* si la durée est supérieure à 30s, et *court* sinon. Le corpus contient 66,0% de segments titrables et 34,0% non titrables respectivement. Cela s'explique par le fait que les JT traitent non seulement de l'actualité du jour mais aussi de sujets de société qui ne se retrouvent pas nécessairement dans les articles de presse du jour.

4.2 Métriques d'évaluation

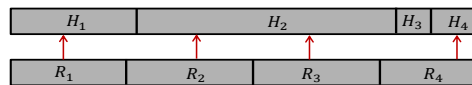
4.2.1 Métrique d'évaluation de la segmentation thématique

La qualité de la ST est généralement évaluée en comparant le positionnement des ruptures thématiques de référence avec celles à évaluer. Les mesures les plus utilisées telles F -mesure, p_k (Beeferman *et al.*, 1999) et $windowdiff$ (Pevzner & Hearst, 2002), donnent une valeur numérique reflétant la performance du système sur la totalité de l'émission. Dans (Boucekif *et al.*, 2014b), une métrique a été proposée pour estimer la performance du système de ST pour chaque segment à détecter. En suivant cette proposition, nous considérons qu'un segment est correct si ses instants de début et fin sont proches de ceux de référence. Pour cela, nous cherchons pour chaque segment de référence R le segment d'hypothèse H correspondant le mieux, c'est à dire le segment ayant une couverture temporelle maximale au segment de référence. La couverture entre deux segments donnés R et H , notée $Couv_{R \leftrightarrow H}$, est définie comme la moyenne harmonique de $Couv_{R \rightarrow H}$ et $Couv_{H \rightarrow R}$ (où $Couv_{R \rightarrow H}$ est le taux de couverture du segment de référence par le segment hypothèse et $Couv_{H \rightarrow R}$ est le taux de couverture du segment hypothèse par le segment de référence). Nous considérons qu'un segment H est correct si $Couv_{R \leftrightarrow H}$ est supérieur à un seuil γ , où

$$Couv_{R \leftrightarrow H} = \frac{2 \times Cov_{R \rightarrow H} \times Cov_{H \rightarrow R}}{Cov_{R \rightarrow H} + Cov_{H \rightarrow R}}. \quad (1)$$

Il ressort de l'exemple de la figure 1 (pour $\gamma = 85\%$) que seul le segment H_1 est correct.

Soit $\#R$ le nombre de segments de référence et $\#H_{Err\gamma}$ le nombre de segments hypothèse dont la couverture harmonique



R_1 correspond à H_1 $Cov_{R_1 \rightarrow H_1} = 100\%$ $Cov_{H_1 \rightarrow R_1} = 86\%$ $Cov_{R_1 \leftrightarrow H_1} = 92\%$
 R_2 correspond à H_2 $Cov_{R_2 \rightarrow H_2} = 94\%$ $Cov_{H_2 \rightarrow R_2} = 48\%$ $Cov_{R_2 \leftrightarrow H_2} = 63\%$
 R_3 correspond à H_2 $Cov_{R_3 \rightarrow H_2} = 100\%$ $Cov_{H_2 \rightarrow R_3} = 45\%$ $Cov_{R_3 \leftrightarrow H_2} = 62\%$
 R_4 correspond à H_4 $Cov_{R_4 \rightarrow H_4} = 38\%$ $Cov_{H_4 \rightarrow R_4} = 100\%$ $Cov_{R_4 \leftrightarrow H_4} = 55\%$

FIGURE 1 – Exemple d'évaluation de la segmentation thématique par nombre de segments corrects.

avec son segment de référence ne dépasse pas le seuil γ . Nous définissons $SE_{R\gamma}$ le taux de segments incorrects tel que

$$SE_{R\gamma} = \frac{\#H_{Err\gamma}}{\#R}. \quad (2)$$

4.2.2 Métrique d'évaluation du titrage

Le système de titrage mis en place peut être évalué selon les erreurs possibles :

- Substitution (*Sub*) : le segment est de type T et le titre affecté au segment n'est pas correct.
- Faux rejet (*FR*) : le segment est de type T et le système ne propose aucun titre.
- Fausse alarme (*FA*) : le segment est de type \bar{T} et le système propose un titre.

Les réponses correctes sont de deux types : (i) le segment est titrable et le titre associé est correct, (ii) le segment est non titrable et aucun titre n'y est associé. TER (Titling Error Rate) est donnée par :

$$TER = \frac{\#Sub + \#FR + \#FA}{\#R} \quad (3)$$

Un titre affecté à un segment est considéré comme bon *si et seulement si* la ST est jugée comme correcte ($Couv_{R \leftrightarrow H} > \gamma$) et le titre affecté figure dans la liste des titres candidats. Pour une évaluation complète, il suffit d'ajouter $H_{Err\gamma}$ comme une source d'erreur.

$$STER_{\gamma} = \frac{\#H_{Err\gamma} + \#Sub_{\gamma} + \#FR_{\gamma} + \#FA_{\gamma}}{\#R} \quad (4)$$

4.3 Expériences et résultats

4.3.1 Évaluation de la segmentation thématique

En terme de nombre de frontières détectées, notre système de ST obtient une F-mesure (de façon similaire à plusieurs travaux dans l'état de l'art, une tolérance de 10s est autorisée entre les frontières d'hypothèses et de références), égale à 71,6% avec un rappel de 73,2% et une précision de 70,0%. Comme il a été mentionné précédemment, nous nous intéressons à l'évaluation en terme de segments corrects. Le tableau 3 résume les résultats obtenus en faisant varier le seuil γ . Nous remarquons une différence de performance entre les segments longs et courts. En effet, contrairement aux segments courts, le système est nettement meilleur sur les segments longs. Ces derniers sont plus importants à extraire que les segments courts (qui ne durent que quelques secondes et correspondent généralement à des brèves).

$SE_{R_{\gamma}}$ (%)	$\gamma = 80\%$	$\gamma = 85\%$	$\gamma = 90\%$
Seg. longs	22,9	26,3	35,1
Seg. courts	72,0	76,7	81,6
Tous les seg.	34,5	38,2	46,0

TABLE 3 – Performance du système en terme de nombre de segments détectés

4.3.2 Évaluation du titrage sur les segments de référence

Dans un premier temps, nous évaluons le titrage sur les segments de référence (définis manuellement) avec différentes mesures de similarité. La performance du système est présentée sous forme d'une courbe ROC (substitution et faux rejet en fonction de fausse alarme) dans la figure 2. Les résultats ont été obtenus en faisant varier le seuil α appliqué sur les valeurs de similarité. La première observation est que les mesures pondérées donnent de meilleurs résultats que le *Jaccard* standard. En effet, la pondération donne plus de poids aux mots importants et pénalise les moins représentatifs, ceci a une influence directe sur le calcul de la cohésion et donc sur le classement des titres candidats. La mesure *Cosine* donne les meilleurs résultats ($Ter = 13,5\%$) avec une légère différence par rapport à *Extended_Jaccard* ($Ter = 13,9\%$). Par la suite, les performances sont données pour la mesure *Cosine*.

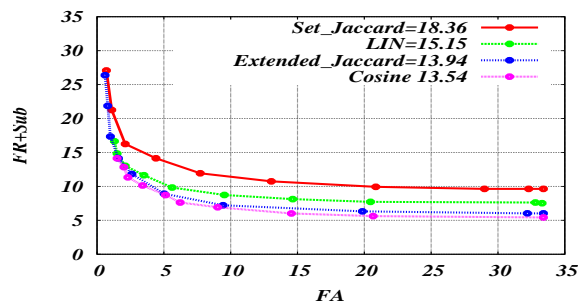


FIGURE 2 – TER pour les différentes variantes de similarité

Le tableau 4 donne la performance du système pour les segments courts et longs. Dans sa globalité, le système donne de très bonnes performances. En effet, entre le titrage et le rejet correct la qualité du titrage est de l'ordre 86,5%. Notre système a la capacité de bien titrer non seulement les segments longs (représentés par un nombre important de termes) mais aussi les segments courts.

Cosine	<i>TER</i>	<i>Sub</i>	<i>FA</i>	<i>FR</i>
Seg. longs	12,6	3,3	2,6	7,6
Seg. courts	16,5	5,1	5,9	5,5
Tous les seg.	13,5	3,7	3,4	6,4

TABLE 4 – Performance du système de titrage sur les segments de référence avec la similarité *Cosine*

Une analyse des erreurs a mis en évidence deux catégories principales d’erreurs : (1) celles qui proviennent de fils d’actualités dont le contenu évolue à court terme, (2) celles qui proviennent d’informations traitant de thématiques proches mais légèrement différentes. Le tableau 5 fournit quelques exemples d’erreurs.

	Titre Correct	Titre erroné	Commentaires
Ex1	Sotchi : la France attend sa première médaille	JO : Martin Fourcade, premier médaillé d’or français.	Dépend de l’heure à laquelle l’information est donnée. Ce genre d’informations peut changer plusieurs fois dans la journée.
Ex2	Le perchiste Renaud Lavillenie accueilli en héros à Roissy.	Saut à la perche : Lavillenie incertain pour les Mondiaux en salle.	
Ex3	Crues : trois départements en vigilance orange.	Tempête : Quelque 10.000 foyers toujours privés d’électricité en Bretagne	Des informations traitent des thématiques proches

TABLE 5 – Exemples d’erreurs de titrage

4.4 Évaluation du titrage sur les segments automatiques

Les résultats du tableau 6 résument le comportement de notre système de titrage sur des segments automatiques. De l’analyse des résultats, il découle que les erreurs de segmentation sont prépondérantes. Plus la qualité de segmentation est bonne plus les erreurs de titrage (*FA*, *Sub*, *FR*) diminuent et donc le système aura la capacité d’affecter de bons titres aux segments. En effet, une faible valeur de couverture signifie que le segment d’hypothèse couvre partiellement le segment de référence (voir le segment H_4 de la figure 1). Par conséquent, la liste des mots décrivant le contenu du segment est incomplète. Dans le cas où le segment couvre partiellement et/ou totalement deux segments de référence (voir le segment H_1 de la figure 1), la liste des mots décrivant le contenu du segment est bruitée.

À noter que les segments courts donnent de mauvaises performances par rapport aux segments longs. En effet, avec une couverture de 85% ($Couv_{R \rightarrow H} > 85$), le $STER_{85}$ global est de 46,8%, le système arrive à segmenter et à titrer correctement 63,6% des segments longs et seulement 19,5% des segments courts.

Pour comparer la robustesse du titrage sur des segments générés manuellement et automatiquement, nous évaluons le titrage sur le même ensemble, pour cela nous écartons les segments mal déterminés ($Couv_{R \rightarrow H} < 85$). Avec les 615 segments corrects, le taux de titres corrects est respectivement de 87% et 86% pour les segments manuels et automatiques. Ces résultats illustrent bien la robustesse de notre système sur les segments thématiques obtenus automatiquement.

γ	$STER_\gamma$	$H_{Err\gamma}$	Sub_γ	FA_γ	FR_γ
80	43,9	34,5	2,4	2,3	4,7
85	46,8	38,2	2,2	2,0	4,4
90	53,3	46,0	1,6	1,8	3,9

TABLE 6 – Taux d’erreur prenant en compte à la fois la segmentation et le titrage.

5 Conclusion

Dans cet article, nous avons décrit notre système de structuration thématique composé de deux tâches complémentaires : la segmentation thématique et le titrage automatique des segments. Après l’étape de ST appliquée à des sorties d’un système de reconnaissance de la parole, les segments obtenus sont titrés. Le titrage consiste à apparier le segment à un article de presse traitant le même sujet. Le titre associé au segment est celui de l’article qui maximise la similarité entre le segment et les articles du jour. Une métrique d’évaluation mesurant conjointement la qualité de la segmentation et du titrage a été proposée. Les résultats obtenus montrent que les erreurs de segmentation restent prépondérantes dans le processus. Le titrage donne de bons résultats et est robuste aux petites imprécisions de la segmentation. Comme perspective à ce travail, il est envisagé d’étudier l’interaction entre ces deux tâches afin d’améliorer la qualité du système de segmentation à partir du titrage, notamment pour les segments courts.

Références

- BEEFERMAN D., BERGER A. L. & LAFFERTY J. D. (1999). Statistical models for text segmentation. *Machine Learning*, **34**(1-3), 177–210.
- BOUCHEKIF A., DAMNATI G. & CHARLET D. (2014a). Intra-content term weighting for topic segmentation. In *39th IEEE International Conference on Acoustics, Speech and Signal Processing, Florence*, p. 7113–7117.
- BOUCHEKIF A., DAMNATI G. & CHARLET D. (2014b). Speech cohesion for topic segmentation of spoken contents. In *INTERSPEECH, Singapore*, p. 1890–1894.
- CHOI F. Y. Y. (2000). Advances in domain independent linear text segmentation. In *Proceedings of the 1st North American Chapter of the Association for Computational Linguistics Conference, NAACL 2000*, p. 26–33.
- CURRAN J. R. & MOENS M. (2002). Improvements in automatic thesaurus extraction. In *Proceedings of the workshop on Workshop On Unsupervised Lexical Acquisition, Philadelphia*, p. 59–66.
- GAUVAIN J.-L., LAMEL L. & ADDA G. (2002). The LIMSI broadcast news transcription system. *Speech Communication*, **37**(1), 89–108.
- HEARST M. A. (1997). Texttiling : Segmenting text into multi-paragraph subtopic passages. *Comput. Linguist.*, **23**(1), 33–64.
- HSUEH P.-Y. & MOORE J. D. (2006). Automatic topic segmentation and labeling in multiparty dialogue. In *Spoken Language Technology Workshop, Aruba*, p. 98–101.
- LAU J. H., GRIESER K., NEWMAN D. & BALDWIN T. (2011). Automatic labelling of topic models. In *Proceedings of the 49th ACL : Human Language Technologies*, p. 1536–1545.
- MALIOUTOV I. & BARZILAY R. (2006). Minimum cut model for spoken lecture segmentation. In *21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference, Sydney*.
- PEVZNER L. & HEARST M. (2002). A critique and improvement of an evaluation metric for text segmentation. *Comput. Linguist.*, **28**(1), 19–36.