Un modèle segmental probabiliste combinant cohésion lexicale et rupture lexicale pour la segmentation thématique

Anca Simon[1] Guillaume Gravier[2] Pascale Sébillot[3]
(1) Université de Rennes 1
(2) CNRS
(3) INSA de Rennes
IRISA & INRIA Rennes
anca-roxana.simon@irisa.fr,guillaume.gravier@irisa.fr,pascale.sebillot@irisa.fr

RÉSUMÉ

L'identification d'une structure thématique dans des données textuelles quelconques est une tâche difficile. La plupart des techniques existantes reposent soit sur la maximisation d'une mesure de cohésion lexicale au sein d'un segment, soit sur la détection de ruptures lexicales. Nous proposons une nouvelle technique combinant ces deux critères de manière à obtenir le meilleur compromis entre cohésion et rupture. Nous définissons un nouveau modèle probabiliste, fondé sur l'approche proposée par Utiyama et Isahara (2001), en préservant les propriétés d'indépendance au domaine et de faible *a priori* de cette dernière. Des évaluations sont menées sur des textes écrits et sur des transcriptions automatiques de la parole à la télévision, transcriptions qui ne respectent pas les normes des textes écrits, ce qui accroît la difficulté. Les résultats expérimentaux obtenus démontrent la pertinence de la combinaison des critères de cohésion et de rupture.

ABSTRACT

A probabilistic segment model combining lexical cohesion and disruption for topic segmentation

Identifying topical structure in any text-like data is a challenging task. Most existing techniques rely either on maximizing a measure of the lexical cohesion or on detecting lexical disruptions. A novel method combining the two criteria so as to obtain the best trade-off between cohesion and disruption is proposed in this paper. A new statistical model is defined, based on the work of Isahara and Utiyama (2001), maintaining the properties of domain independence and limited a priori of the latter. Evaluations are performed both on written texts and on automatic transcripts of TV shows, the latter not respecting the norms of written texts, thus increasing the difficulty of the task. Experimental results demonstrate the relevance of combining lexical cohesion and disrupture.

MOTS-CLÉS : segmentation thématique, cohésion lexicale, rupture de cohésion, journaux télévisés.

KEYWORDS: topic segmentation, lexical cohesion, lexical disrupture, TV broadcast news.

1 Introduction

La segmentation thématique consiste à mettre en évidence la structure sémantique d'un document et les algorithmes développés pour cette tâche visent à détecter automatiquement les frontières qui définissent des segments thématiquement cohérents. Cible de nombreux travaux, la segmentation thématique a également des retombées en recherche d'information, résumé automatique, systèmes de question-réponse...

Diverses méthodes de segmentation de données textuelles ont été proposées dans la littérature (Yamron et al., 1998; Georgescul et al., 2006; Galley et al., 2003; Hearst, 1997; Reynar, 1994; Moens and Busser, 2001; Choi, 2000; Ferret et al., 1998; Utiyama and Isahara, 2001). Comme indiqué dans (Purver, 2011), elles peuvent être supervisées ou non, reposer sur des changements de vocabulaire, des techniques de *clustering*, sur la détection de frontières discriminantes ou sur des modèles probabilistes. Déterminer les segments thématiques à l'aide de modèles probabiliste consiste la plupart du temps à inférer la séquence de thèmes la plus probable à partir des mots observés et à dériver les positions des frontières (Yamron et al., 1998; Blei and Moreno, 2001). Ces modèles utilisent un corpus d'apprentissage pour estimer les distributions documents-thèmes et thèmes-mots. Des travaux récents ont montré l'intérêt de l'intégration de ces modèles probabilistes dans les algorithmes de segmentation de textes reposant sur la similarité de vocabulaire (Misra and Yvon, 2010; Riedl and Biemann, 2012). Nos travaux portent sur les méthodes non supervisées. La plupart d'entre elles repose sur la cohésion du vocabulaire pour identifier des segments cohérents dans les textes, exploitant les mots qu'ils contiennent et les relations sémantiques que ces mots entretiennent. Pour mesurer la cohérence dans les (segments de) textes, la cohésion lexicale, fondée sur la répétition de mots ou sur l'exploitation de chaînes lexicales, est fréquemment retenue en privilégiant l'une ou l'autre des deux stratégies suivantes : soit on cherche à maximiser la mesure de cohésion lexicale des segments, en regroupant les portions de texte lexicalement cohérentes, soit on cherche à identifier des ruptures entre les segments en plaçant des frontières quand survient un changement significatif dans le vocabulaire utilisé (Hearst, 1997). Dans cet article, notre objectif est de proposer une nouvelle solution pour la segmentation thématique de documents qui consiste à mêler ces deux approches, c'est-à-dire à combiner les mesures de *cohésion lexicale* et de *rupture lexicale* afin d'obtenir une segmentation en fragments à la fois thématiquement cohérents et différents les uns des autres.

La technique que nous proposons peut s'appliquer à tout type de données textuelles et est indépendante d'un domaine particulier. Notre objectif est cependant de l'appliquer à la segmentation de journaux télévisés afin de permettre à des utilisateurs de naviguer dans ce type de données. De manière à rester générique et non supervisée, la segmentation thématique peut dans ce cas s'appuyer sur la transcription automatique de la parole prononcée dans les émissions. L'analyse des mots de la transcription vise alors à trouver un changement significatif de vocabulaire et donc un changement de thème (Hearst, 1997). Cependant, les particularités des transcriptions automatiques accroissent la difficulté de la tâche de segmentation. En effet, ces transcriptions ne contiennent ni casse, ni ponctuation, et ne sont donc pas structurées en phrases comme des textes standards mais en groupes de soufle correspondant aux mots prononcés par une personne entre deux inspirations. De plus, elles peuvent contenir de nombreux mots mal transcrits. Difficulté supplémentaire, les journaux TV peuvent avoir des segments thématiques très courts, contenant peu de mots et donc peu de répétitions, en particulier quand le présentateur fait volontairement usage de synonymes. Cela rend l'utilisation du critère de cohésion lexicale particulièrement ardue. Notre algorithme de segmentation thématique ayant un fort potentiel pour traiter ces cas, nous

avons souhaité le tester sur ces données difficiles.

La technique présentée ici repose sur l'algorithme de segmentation de textes proposé par Utiyama et Isahara (2001), algorithme dont les capactiés ont été attestées pour le texte écrit. C'est un modèle probabiliste qui fournit une segmentation non supervisée. Dans cette approche, il n'y a donc pas de tentative d'apprentissage de l'ensemble des modèles thématiques le plus probable à partir des données d'apprentissage, mais au contraire l'ensemble est généré par l'algorithme étant donnés les textes à segmenter. Ce modèle est indépendant du domaine et permet l'obtention de segments de longueurs très variées. Il consiste en une représentation du document à segmenter sous forme d'un graphe, où les nœuds représentent les frontières thématiques potentielles et les arcs les segments. La segmentation thématique est obtenue en trouvant le meilleur chemin dans le graphe valué, dans lequel les poids reflètent la valeur de cohésion lexicale. Notre contribution consiste à définir un modèle statistique amélioré qui permet d'intégrer la rupture lexicale. Par conséquent, notre algorithme se résume en un décodage d'un treillis afin d'identifier la meilleure segmentation. Cette représentation permet de considérer la valeur de rupture lexicale en chaque nœud. La solution proposée est testée pour la segmentation de journaux TV transcrits mais également de textes écrits, et les évaluations montrent une amélioration en précision et rappel par rapport à la seule utilisation de la valeur de la cohésion lexicale.

L'article est organisé de la façon suivante : des travaux en segmentation thématique existants sont présentés dans la section 2. Dans la section 3, nous détaillons notre approche, en décrivant d'abord le modèle général d'Utiyama et Isahara puis le nouveau modèle statistique proposé. Dans la section 4, les expériences sont présentées, avec des détails sur les corpus utilisés et une analyse des résultats.

2 Techniques de segmentation thématique

Dans cette section, nous présentons rapidement les notions-clés concernant le concept de segmentation thématique, ainsi que les techniques existantes et les traits qu'elles exploitent pour réaliser cette tâche.

2.1 Le concept de thème

Le concept de thème est difficile à définir précisément et les linguistes qui ont tenté de le caractériser en offrent de nombreuses définitions. Dans (Brown and Yule, 1983), la difficulté de définir un thème est longuement discutée et les auteurs soulignent que : "*The notion of 'topic' is clearly an intuitively satisfactory way of describing the unifying principle which makes one stretch of discourse 'about' something and the next stretch 'about' something else, for it is appealed to very frequently in the discourse analysis literature. Yet the basis for the identification of 'topic' is rarely made explicit.*".

Souhaitant appliquer la segmentation thématique à des journaux TV, nous avons cherché à voir si la notion de thème avait été définie dans le contexte d'émissions télévisées. Le projet *Topic Detection and Tracking* (Allan, 2002) s'est par exemple focalisé sur le repérage de segments de journaux TV thématiquement liés. Dans ce cadre, les notions d'événement et de thème ont été définies : un événement est quelque chose qui se produit à un instant et un endroit spécifique

et qui est associé à des actions particulières ; un thème est, quant à lui, l'ensemble formé d'un événement et de tous les événements qui lui sont directement liés. Un événement est donc relativement court et évolue dans le temps, tandis qu'un thème est plus stable et plus long.

Dans notre cadre de segmentation de journaux TV, un thème correspond à un reportage qui forme une unité sémantique cohérente dans la structure d'un journal. Notre algorithme est également évalué sur des textes écrits, formés par concaténation de parties extraites d'articles sélectionnés aléatoirement dans le corpus Brown (Choi, 2000) ; un thème est alors associé à chaque partie formant le texte final.

2.2 Méthodes pour la segmentation thématique

Pour réaliser la segmentation thématique de textes, diverses caractéristiques peuvent être exploitées afin d'identifier les changements thématiques. Elles peuvent reposer sur la cohésion lexicale (*i.e.*, prendre en compte les informations de distribution du vocabulaire) ou sur des marqueurs linguistiques tels que des indices prosodiques (Guinaudeau and Hirschberg, 2011) ou des marqueurs du discours (Grosz and Sidner, 1986; Litman and Passonneau, 1995). Les techniques génériques, qui sont celles qui nous intéressent ici, exploitent traditionnellement la seule cohésion lexicale, indépendante du type de documents considérés et ne nécessitant pas de phase d'apprentissage. L'idée-clé des méthodes fondées sur la cohésion lexicale est de considérer qu'un changement significatif dans le vocabulaire utilisé est un signe de changement thématique. Ces approches peuvent être divisées en deux familles :

- les méthodes locales (Hearst, 1997; Hernandez and Grau, 2002; Ferret et al., 1998; Claveau and Lefèvre, 2011) qui cherchent à repérer localement les *ruptures lexicales* ;
- les méthodes globales (Reynar, 1994; Choi, 2000; Utiyama and Isahara, 2001; Malioutov and Barzilay, 2006; Misra and Yvon, 2010) exploitant une *mesure de la cohésion lexicale*.

Une méthode locale repose sur la comparaison locale de régions du document et associe un changement thématique aux endroits où il y a une similarité faible entre deux régions consécutives (*i.e.*, elles identifient les zones de fortes ruptures lexicales). Par exemple, TextTiling (Hearst, 1997), qui est considéré comme un algorithme de segmentation thématique fondamental, analyse le texte à l'aide d'une fenêtre glissante qui couvre des blocs adjacents de texte et est centrée en un point du texte correspondant à une frontière thématique potentielle. Les contenus avant et après chaque frontière possible sont représentés par des vecteurs de mots pondérés, un poids fort indiquant qu'un mot est particulièrement pertinent pour décrire un contenu. Une mesure de similarité, par exemple cosinus, est calculée entre les deux vecteurs. Plus l'angle entre les deux vecteurs diminue, plus le cosinus approche de 1, indiquant par là-même la plus grande similarité entre les contenus avant et après la frontière potentielle. Les valeurs de similarité sont calculées à chaque frontière possible et la séquence résultante de valeurs de similarité est analysée. Les points de scores de similarité les plus bas (*i.e.*, forte rupture) représentent alors les frontières thématiques. Ce type de méthode locale présente certains désavantages dont une sensibilité aux variations de tailles des segments dans les textes puisqu'un voisinage de taille fixe est considéré, ainsi qu'une difficulté de choix de la valeur de seuil pour décider qu'une rupture est suffisamment forte pour placer une frontière.

Une méthode globale réalise quant à elle une comparaison globale entre toutes les régions du document, en cherchant à maximiser globalement la valeur de la cohésion lexicale. Dans Utiyama et Isahara (2001), 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, appris sur le segment S_i, à prédire les mots du segment. Le modèle de langue Δ_i doit donc d'abord être estimé, puis la probabilité généralisée des mots du segment S_i, étant donné Δ_i, doit être déterminée. Après le calcul de la valeur de cohésion lexicale pour chaque segment, la segmentation maximisant globalement cette valeur est choisie. Cet algorithme s'est avéré performant au regard d'autres algorithmes de segmentation thématique de textes tels que ceux de Choi (2000) ou Reynar (1994). Cependant, la limite principale de ce type de méthode globale est un risque de sur-segmentation.

L'originalité de la solution que nous proposons consiste dans la combinaison des deux types de méthodes. Une méthode fondée sur le même principe, visant à capturer dans une vue globale des dissimilarités locales, a été présentée dans (Malioutov and Barzilay, 2006), mais, d'une part, le nombre de segments à trouver est fixé *a priori* et, d'autre part, la couverture est limitée car la dissimilarité entre segments est calculée en utilisant une fenêtre.

Le point de départ de notre méthode est le modèle statistique proposé dans (Utiyama and Isahara, 2001), qui est flexible et offre des possibilités d'extension par intégration de nouvelles informations. Plusieurs travaux l'ont déjà utilisé avec succès dans le contexte de la segmentation de journaux TV (Huet et al., 2008; Guinaudeau et al., 2012), le modifiant pour intégrer des connaissances spécifiques aux émissions TV. Contrairement à ces travaux, nous avons redéfini le modèle de (Utiyama and Isahara, 2001) afin qu'il puisse prendre en compte non seulement la cohésion mais aussi la rupture lexicale et, par conséquent, améliorer la segmentation de tout type de données textuelles. Considérer la rupture est en particulier intéressant pour traiter les cas de textes contenant des changements brutaux de vocabulaire. La façon dont nous combinons les deux critères est détaillée dans la section 3.

3 Combinaison de la cohésion et de la rupture lexicales

Nous rappelons tout d'abord l'algorithme de Utiyama et Isahara, puis expliquons le nouveau modèle statistique que nous proposons.

3.1 Le modèle statistique

L'algorithme proposé par Utiyama et Isahara définit un modèle probabiliste et consiste à déterminer 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. L'idée principale est de trouver la segmentation la plus probable pour une séquence de t unités élémentaires (*i.e.*, phrases ou énoncés composés de mots) $W = u_1^t$ parmi toutes les segmentations possibles, *i.e.*,

$$\hat{S} = \arg\max_S P[W|S]P[S] \ . \tag{1}$$

En admettant que chaque segment est une unité indépendante du reste du texte et que les mots contenus dans un segment sont eux aussi indépendants, la probabilité du texte W pour une segmentation $S = S_1^m$ est donnée par

$$P[W|S_1^m] = \prod_{i=1}^{m} \prod_{j=1}^{n_i} P[w_j^i|S_i] \ , \tag{2}$$

où n_i est le nombre de mots du segment S_i, w_j^i est le j^e mot de S_i et m le nombre de segments. La probabilité $P[w_j^i|S_i]$ est donnée par une loi de Laplace dont les paramètres sont estimés sur S_i, *i.e.*,

$$P[w_j^i|S_i] = \frac{f_i(w_j^i)+1}{n_i+k} \ , \tag{3}$$

où $f_i(w_j^i)$ est le nombre d'occurrences de w_j^i dans S_i et k est le nombre total de mots différents dans le texte W (*i.e.*, la taille du vocabulaire). Cette probabilité va favoriser les segments homogènes car elle croît quand les mots sont répétés et décroît quand ils sont différents. La distribution *a priori* des longueurs des segments est donnée par $P[S_1^m] = n^{-m}$, où n est le nombre total de mots. Elle a une valeur élevée quand le nombre de segments est faible, tandis que $P[W|S]$ a des valeurs élevées quand le nombre de segments est grand.

Cette approche peut être vue comme la recherche du meilleur chemin dans un graphe valué, graphe représentant toutes les segmentations possibles. Chaque nœud correspond à une frontière possible et un arc entre les nœuds i et j représente un segment contenant les unités comprises entre u_{i+1} et u_j. Le poids attribué à chaque arc de ce type est

$$v(i,j) = \sum_{k=i+1}^{j} ln(P[u_k|S_{i\to j}]) - \alpha ln(n) \ , \tag{4}$$

où $S_{i\to j}$ est le segment correspondant à l'arc allant du nœud i au nœud j. Pour les petits segments, la probabilité d'estimer les mots contenus dans le segment est plus faible ; le facteur α fournit un compromis entre la longueur moyenne des segments retournés et la valeur de la cohésion lexicale.

3.2 Introduction de la rupture lexicale

Le modèle défini ci-dessus suppose que chaque segment S_i du texte est indépendant des autres, ce qui ne permet pas de combiner la valeur de la cohésion lexicale et celle de la rupture lexicale. En effet, lors du calcul du poids associé au segment S_i, nous devrions ajouter une pénalité marquant à quel point le contenu de S_i diffère de celui du segment précédent S_{i-1}. Pour cette raison, nous proposons une hypothèse markovienne entre les segments nous permettant, pour chaque segment, de considérer celui qui le précède. La probabilité d'un texte W pour une segmentation $S = S_1^m$ devient alors

$$P[W|S_1^m] = P[W|S_1] \prod_{i=2}^{m} P[W|S_i, S_{i-1}] \ . \tag{5}$$

Pour déterminer la segmentation de probabilité maximum $\hat{S}$, le coût associé au segment S_i, étant donné S_{i-1}, est

$$ln(P[W|S_i, S_{i-1}]) = ln(P[W_i|S_i]) - \lambda(\frac{1}{\Delta(W_i, W_{i-1})}) \ , \tag{6}$$

où $\Delta(W_i, W_{i-1})$ est la valeur de rupture entre le contenu de S_i et celui de S_{i-1}, et λ est un paramètre qui permet de contrôler l'influence de la rupture dans le coût. W_i représente les unités élémentaires du segment S_i. Choisir $1/\Delta(W_i, W_{i-1})$ conduit à une pénalité faible quand il y a une forte rupture. Dans l'équation 6, $P[W|S_i, S_{i-1}]$ ne représente plus une probabilité ; cependant,

puisque l'algorithme de segmentation consiste à déterminer le meilleur chemin dans un graphe pondéré, cela n'a pas d'impact car aucune présupposition de graphe probabiliste n'est faite pour segmenter. Par conséquent, la nouvelle définition de la segmentation la plus probable est

$$\hat{S} = \arg\max_{S} \sum_{i=1}^{m} ln(P[W_i|S_i]) - \lambda \sum_{i=2}^{m} (\frac{1}{\Delta(W_i, W_{i-1})}) - \alpha ml n(n) \ . \quad (7)$$

De l'équation 6, on peut déduire que, pour un nœud donné représentant une frontière thématique, tous les segments de longueurs différentes arrivant à ce nœud sont conservés. Au niveau implémentation, nous définission un treillis dans lequel un arc $e_{ip,jl}$ représente une prolongation d'un chemin de longueur l du nœud n_{ip} au nœud n_{jl}. Un nœud n_{ip} rassemble donc tous les segments de longueur p unités se terminant après u_i. Ceci signifie qu'en chaque point du texte où une frontière potentielle est considérée, nous analysons toutes les combinaisons possibles d'unités consécutives précédant cette frontière. Un arc $e_{ip,jl}$ représente un segment contenant toutes les unités entre u_{i+1} et u_j, avec $j - i = l$. Un coût est associé à chaque arc en se fondant sur l'équation 6. D'une part, ce coût consiste en la valeur de la cohésion lexicale du segment couvert par l'arc calculé grâce à l'équation 3 ; d'autre part, une pénalité est associée à chacune des valeurs de ce type, en fonction de la rupture lexicale entre le segment couvert par l'arc et le segment précédent dans le texte. Selon le nœud dont il provient, le segment précédent peut lui aussi avoir différentes longueurs. Par conséquent, la rupture est calculée entre toutes les paires possibles de segments. Pour obtenir la rupture, une mesure de similarité cosinus est utilisée entre les vecteurs représentant

- le segment qui contient les unités couvertes par l'arc (de score le plus élevé) arrivant au nœud $n_{i,j}$ et
- le segment qui contient les unités couvertes par l'arc sortant de ce nœud vers $n_{i+k,k}$.

Les vecteurs contiennent les poids associés aux mots dans les unités. Ces poids sont calculés en utilisant les mesures de TF-IDF et Okapi (Claveau, 2012), transformées en dissimilarités.

Pour déterminer la meilleure segmentation, nous utilisons un algorithme de programmation dynamique. Lors du décodage, on associe à chaque nœud le coût du meilleur chemin en fonction des arcs entrants. Par exemple dans la figure 1, les calculs au nœud $n_{3,1}$ consistent à choisir la valeur la plus élevée entre le poids associé à l'arc $e_{21,31}$ et à l'arc $e_{22,31}$.

- Pour le premier arc, le score est donné par la valeur associée au nœud $n_{2,1}$, la valeur de la cohésion lexicale de l'arc $e_{21,31}$ et la rupture entre le segment contenant u_2 et le segment contenant u_3.
- Pour le second, le score est donné par la valeur associée au nœud $n_{2,2}$, la cohésion lexicale de l'arc $e_{22,31}$ et la rupture lexicale entre le segment contenant à la fois u_1 et u_2 et le segment contenant seulement u_3.

Si dans l'exemple donné (*cf.* FIGURE 1) le score le plus élevé est obtenu pour le chemin formé de $e_{01,11}e_{11,32}e_{32,41}$, la segmentation de probabilité maximum est $[u_1][u_2u_3][u_4]$. Utiliser cette représentation nous permet donc de considérer tous les chemins possibles de longueurs variables, traitant ainsi toutes les combinaisons possibles de segments consécutifs pour le calcul de la cohésion lexicale et également de la rupture lexicale.

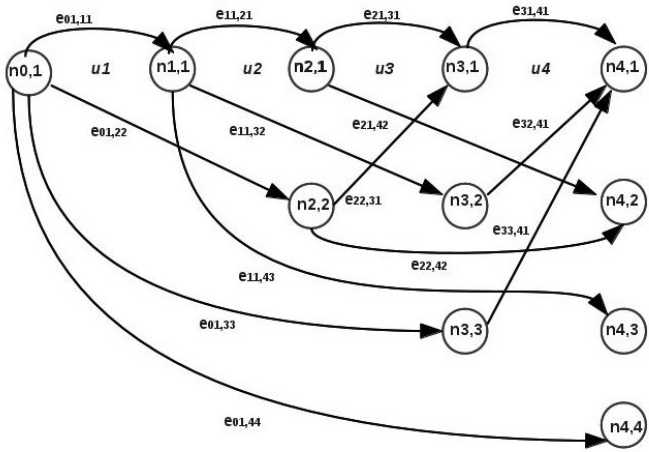


FIGURE 1 – Un exemple de treillis de segmentation

4 Expériences

Nous présentons ici les expériences réalisées en fournissant tout d'abord des détails sur les transcriptions de journaux TV et les données textuelles utilisées, puis en analysant les résultats obtenus.

4.1 Corpus

Deux corpora sont considérés dans notre tâche de segmentation thématique. Le premier est un corpus de journaux TV contenant 56 journaux (∼1/2 heure chacun), enregistrés de février à mars 2007 sur la chaîne de TV française France 2. Les journaux consistent en une succession de reportages de courte durée (2-3 mn), contenant très peu de répétitions de mots par rapport à d'autres types d'émissions, des synonymes étant fréquemment préférés. Les transcriptions utilisées dans les expériences proviennent de deux systèmes de transcription : IRENE, le système de l'IRISA, et LIMSI, le système du Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Ingénieur. IRENE a un taux d'erreurs mots plus élevé d'environ 7%. La segmentation de référence a été créée en associant un thème à chaque reportage. Les frontières thématiques sont donc placées au début de l'introduction du reportage et à la fin de ses remarques conclusives.

Le second corpus est un jeu de données artificiel proposé par Choi (2000) et utilisé par différents auteurs pour comparer leurs méthodes à des approches existantes. Il consiste en 700 documents créés par concaténation de 10 parties de textes correspondant chacune aux z premières phrases d'articles choisis aléatoirement dans le corpus Brown, z étant lui-même choisi aléatoirement dans un intervalle fixé. Une limite de ce jeu de données est qu'il comporte donc des changements thématiques très brutaux, ce qui est rarement le cas dans des documents classiques. Cependant, il est intéressant car il contient des segments de longueurs variables.

Transcriptions	Manuelles	IRENE automatiques	LIMSI automatiques
Gain de F1-mesure	0.77	0.2	0.5

TABLE 1 – Gain en F1-mesure pour les transcriptions manuelles et automatiques de journaux TV

4.2 Résultats

Nous présentons dans cette sous-section l'impact de notre modèle statistique sur la tâche de segmentation thématique de journaux TV et de données textuelles. Les résultats sont comparés à ceux d'un système basique et bien que les améliorations obtenues soient limitées, elles montrent nettement l'intérêt de combiner rupture et cohésion lexicales. Pour les journaux TV, le traitement de ces données difficiles diminue les capacités de notre méthode et, pour cette raison, des transcriptions manuelles ont également été considérées lors des expériences.

Pour l'évaluation, des mesures de rappel, précision et F1-mesure ont été utilisées après alignement de la référence et des frontières proposées. Une tolérance de 10 secondes dans le positionnement est autorisée dans le cas des transcriptions de journaux TV, et de 2 phrases pour les données textuelles. Le rappel correspond à la part de frontières de référence détectées par la méthode et la précision au ratio des frontières produites appartenant à la segmentation de référence. La F1-mesure combine rappel et précision en une valeur unique. D'autres mesures ont été précédemment proposées pour évaluer la segmentation thématique de textes. Cependant, contrairement à la mesure *Pk* (Beeferman et al., 1997), le rappel et la précision ne sont pas sensibles aux variations de tailles des segments et ces mesures ne favorisent pas les segmentations avec peu de frontières comme la mesure *WindowDiff* (Pevzner and Hearst, 2002), ce qui justifie notre choix.

Les tests effectués ont consisté à faire varier les paramètres α et λ de l'équation 7, α permettant différents compromis entre les valeurs de précision et de rappel, tandis que λ donne plus ou moins d'importance à la rupture.

Parmi les diverses configurations testées dans les expériences, seules quelques-unes sont présentées ici. La figure 2 illustre tout d'abord les résultats obtenus pour la segmentation des journaux TV transcrits par les deux systèmes de RAP, en les comparant au système de référence correspondant à l'algorithme d'Utiyama et Isahara (2001) standard. Les valeurs présentées correspondent à des pondérations TF-IDF lors de l'évaluation de la rupture lexicale, les résultats obtenus avec Okapi étant similaires. Nous constatons que les précision et rappel pour le corpus LIMSI sont supérieurs à ceux du corpus IRENE, ce qui se justifie par le taux d'erreur de transcription plus élevé de ce dernier. Notre méthode reposant sur la cohérence du vocabulaire, l'amélioration assez faible obtenue par rapport au système étalon s'explique par le fait que les transcriptions sont des données difficiles, contenant des segments très courts et peu de répétitions. Le gain en F1-mesure lors de la segmentation des transcriptions manuelles et automatiques est donné dans le tableau 1. Ces résultats ne concernent toutefois que 6 journaux TV, la F1-mesure retenue correspondant aux segmentations fournissant le nombre de frontières le plus proche de celui de la référence. Le gain est inférieur là encore pour les transcriptions IRENE dont le taux d'erreur est plus élevé. Avoir à sa disposition moins de mots potentiellement répétés accroît la difficulté de discriminer entre des segments appartenant à des thèmes différents. Cependant notre modèle parvient à améliorer la segmentation même pour ces données bruitées.

Notre méthode offrant une amélioration limitée sur la segmentation des transcriptions de

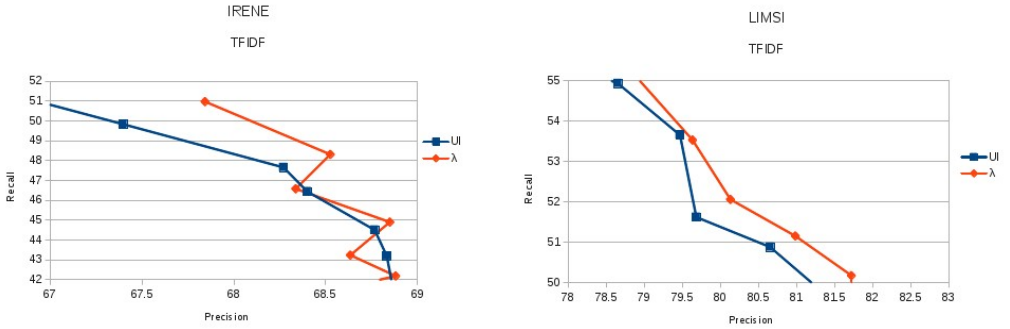
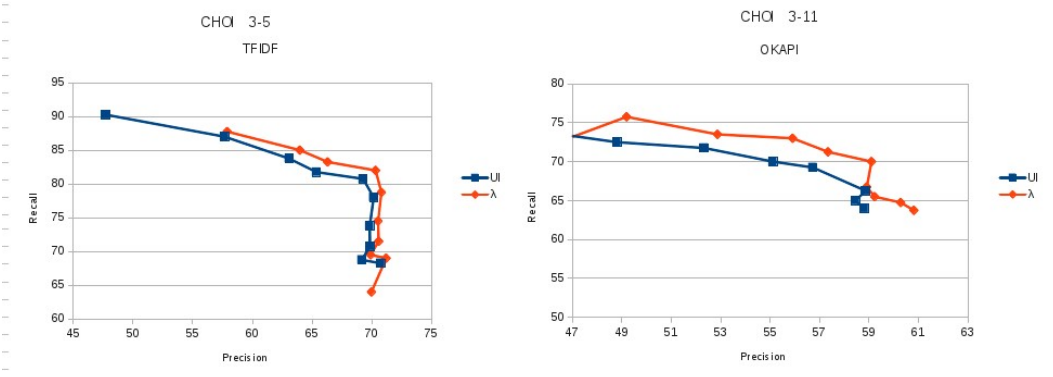


FIGURE 2 – Courbe rappel/précision pour les transcriptions obtenues grâce aux systèmes de reconnaissance de la parole LIMSI et IRENE. UI représente les résultats obtenus grâce à la seule cohésion lexicale ; $\lambda - value$ indique l'importance donnée à la rupture lexicale dans notre approche

journaux TV, nous avons également utilisé le corpus de Choi afin de vérifier que notre modèle fonctionnait bien sur des données plus classiques. Par ailleurs, le jeu de données artificiel de Choi nous permet d'observer le comportement de notre approche lorsque les longueurs des segments varient. Les résultats de notre méthode sur le corpus de Choi sont présentés sur la figure 3.

Les nombres mentionnés sur chaque figure (par exemple 3-5, 3-11) correspondent à l'intervalle de valeurs pour z. Les résultats de différents échantillons du jeu de données sont fournis. On observe que lorsque notre algorithme traite des textes écrits, il obtient de meilleures performances, augmentant les valeurs de rappel et de précision. Plus les segments sont longs en moyenne, plus importante est l'amélioration apportée par la prise en compte de la rupture. Cependant les paramètres utilisés doivent encore être ajustés pour que l'importance donnée à la rupture, pour tout type de données, soit fixée et soit capable d'assigner la pénalité nécessaire aux poids calculés. Nous avons observé qu'il ne semble pas y avoir de valeur précise à donner à l'importance de la rupture ; cependant les valeurs plus élevées conduisent à un rappel plus bas et une précision plus élevée, conduisant à une sous-segmentation.



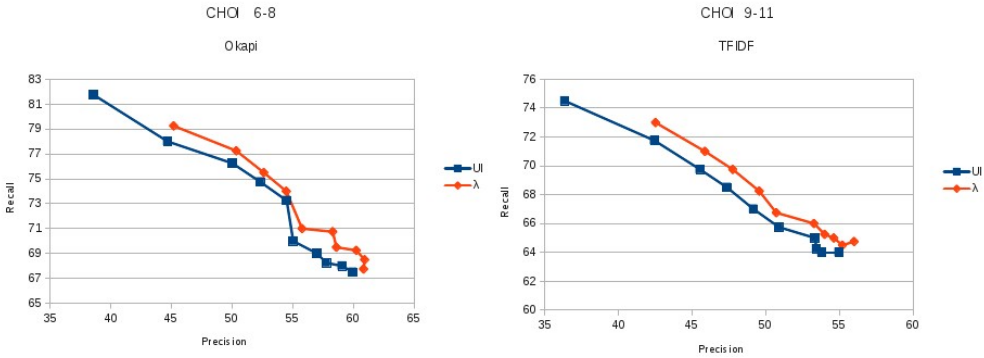


FIGURE 3 – Courbes rappel/précision obtenues sur le corpus de Choi

5 Conclusions

Nous avons proposé une méthode originale de segmentation thématique qui combine la cohésion lexicale et la rupture lexicale, identifiant des zones de continuités et de ruptures dans l'organisation globale des données. Les résultats obtenus montrent que la combinaison des deux mesures produit des segmentations de meilleure qualité que lors de l'emploi de la seule cohésion lexicale. Il reste toutefois encore des possibilités d'améliorer notre approche.

Nous proposons comme perspectives d'employer d'autres techniques de calcul de la rupture lexicale. Parmi elles, la vectorisation (Claveau and Lefèvre, 2011) implique une comparaison indirecte entre des segments consécutifs, en proposant un changement dans l'espace de représentation des segments et l'utilisation de documents pivots pour le calcul de la rupture. Les segments ne partageant pas beaucoup de vocabulaire quoiqu'abordant le même thème pourraient alors être considérés comme similaires. Cette méthode pourrait donc permettre de pallier le manque de répétitions de mots qui apparaît particulièrement dans le cas de transcriptions de journaux TV. Par ailleurs, une façon de régler finement les paramètres α and λ utilisés dans notre modèle statistiques doit être déterminée.

Références

Allan, J., editor (2002). *Topic Detection and Tracking : event-based information organization*. Kluwer Academic Publishers.

Beeferman, D., Berger, A., and Lafferty, J. (1997). Text segmentation using exponential models. *In Proceedings of the 2nd Conference on Empirical Methods in Natural Language Processing*, pages 35–46.

Blei, D. and Moreno, P. (2001). Topic segmentation with an aspect hidden Markov model. *In Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 343–348.

Brown, G. and Yule, G. (1983). Discourse analysis. *Cambridge University Press*.

Choi, F. Y. Y. (2000). Advances in domain independent linear text segmentation. *In Proceedings of the 1st International Conference of the North American Chapter of the Association for Computational Linguistics*, pages 26–33.

Claveau, V. (2012). Vectorisation, Okapi et calcul de similarité pour le TAL : pour oublier enfin le TF-IDF. *Actes de la conférence conjointe JEP-TALN-RECITAL 2012*, pages 85–98.

Claveau, V. and Lefèvre, S. (2011). Topic segmentation of TV-streams by mathematical morphology and vectorization. *In Proceedings of the 12th International Conference of the International Speech Communication Association, Interspeech'11*, pages 1105–1108.

Ferret, O., Grau, B., and Masson, N. (1998). Thematic segmentation of texts : Two methods for two kinds of texts. *In Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics*, pages 392–396.

Galley, M., McKeown, K., Fosler-Lussier, E., and Jing, H. (2003). Discourse segmentation of multi-party conversation. *In Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics, ACL*, pages 562–569.

Georgescul, M., Clark, A., and Armstrong, S. (2006). Word distributions for thematic segmentation in a support vector machine approach. *In Proceedings of the 10th Conference on Computational Natural Language Learning, CoNLL-X*, pages 101–108.

Grosz, B. J. and Sidner, C. L. (1986). Attention, intentions, and the structure of discourse. *Computational Linguistics*, 12(3) :175–204.

Guinaudeau, C., Gravier, G., and Sébillot, P. (2012). Enhancing lexical cohesion measure with confidence measures, semantic relations and language model interpolation for multimedia spoken content topic segmentation. *Computer Speech and Language*, 26(2) :90–104.

Guinaudeau, C. and Hirschberg, J. (2011). Accounting for prosodic information to improve ASR-based topic tracking for TV broadcast news. *In 12th Annual Conference of the International Speech Communication Association, Interspeech'11*, pages 1401–1404.

Hearst, M. A. (1997). TextTiling : Segmenting text into multi-paragraph subtopic passages. *Computational Linguistics*, 23(1) :33–64.

Hernandez, N. and Grau, B. (2002). Analyse thématique du discours : segmentation, structuration, description et représentation. *In Actes du 5e colloque international sur le document électronique*, pages 277–285.

Huet, S., Gravier, G., and Sébillot, P. (2008). Un modèle multi-sources pour la segmentation en sujets de journaux radiophoniques. In *Actes de 15e conférence sur le traitement automatique des langues naturelles, TALN'08*, pages 49–58.

Litman, D. J. and Passonneau, R. J. (1995). Combining multiple knowledge sources for discourse segmentation. *In Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics*, pages 108–115.

Malioutov, I. and Barzilay, R. (2006). Minimum cut model for spoken lecture segmentation. *In Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics*, pages 25–32.

Misra, H. and Yvon, F. (2010). Modèles thématiques pour la segmentation de documents. *In Actes des 10e journées internationales d'analyse statistique des données textuelles*, pages 203–213.

Moens, M.-F. and Busser, R. D. (2001). Generic topic segmentation of document texts. *In Proceedings of the 24th International Conference on Research and Developement in Information Retrieval*, pages 418–419.

Pevzner, L. and Hearst, M. A. (2002). A critique and improvement of an evaluation metric for text segmentation. *Computational Linguistics*, 28 :19–36.

Purver, M. (2011). Topic segmentation. In Tur, G. and de Mori, R., editors, *Spoken Language Understanding : Systems for Extracting Semantic Information from Speech*, chapter 11, pages 291–317. Wiley.

Reynar, J. C. (1994). An automatic method of finding topic boundaries. *In Proceedings of the 32nd Annual Meeting on Association for Computational Linguistics*, pages 331–333.

Riedl, M. and Biemann, C. (2012). How text segmentation algorithms gain from topic models. *In Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, pages 553–557.

Utiyama, M. and Isahara, H. (2001). A statistical model for domain-independent text segmentation. *In Proceedings of the 39th Annual Meeting on the Association for Computational Linguistics*, pages 499–506.

Yamron, J., Carp, I., Gillick, L., Lowe, S., and van Mulbregt P. (1998). A hidden Markov model approach to text segmentation and event tracking. *In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP*, pages 333–336.