Évaluation d'une nouvelle structuration thématique hiérarchique des textes dans un cadre de résumé automatique et de détection d'ancres au sein de vidéos

Anca Simon[1] Guillaume Gravier[2] Pascale Sébillot[3]
(1) Université de Rennes 1, IRISA & INRIA Rennes, Campus de Beaulieu, 35042 Rennes, France
(2) CNRS, IRISA & INRIA Rennes, Campus de Beaulieu, 35042 Rennes, France
(3) INSA, IRISA & INRIA Rennes, Campus de Beaulieu, 35042 Rennes, France
anca.simon@irisa.fr, guillaume.gravier@irisa.fr,
pascale.sebillot@irisa.fr

RÉSUMÉ

Dans cet article, nous évaluons, à travers son intérêt pour le résumé automatique et la détection d'ancres dans des vidéos, le potentiel d'une nouvelle structure thématique extraite de données textuelles, composée d'une hiérarchie de fragments thématiquement focalisés. Cette structure est produite par un algorithme exploitant les distributions temporelles d'apparition des mots dans les textes en se fondant sur une analyse de salves lexicales. La hiérarchie obtenue a pour objet de filtrer le contenu non crucial et de ne conserver que l'information saillante des textes, à différents niveaux de détail. Nous montrons qu'elle permet d'améliorer la production de résumés ou au moins de maintenir les résultats de l'état de l'art, tandis que pour la détection d'ancres, elle nous conduit à la meilleure précision dans le contexte de la tâche *Search and Anchoring in Video Archives* à MediaEval. Les expériences sont réalisées sur du texte écrit et sur un corpus de transcriptions automatiques d'émissions de télévision.

ABSTRACT

Evaluation of a novel hierarchical thematic structuring of texts in the framework of text summarization and anchor detection for video hyperlinking

This paper investigates the potential of a novel topical structure of text-like data in the context of summarization and anchor detection in video hyperlinking. This structure is produced by an algorithm that exploits temporal distributions of words through word burst analysis to generate a hierarchy of topically focused fragments. The obtained hierarchy aims at filtering out non-critical content, retaining only the salient information at various levels of detail. For the tasks we choose to evaluate the structure on, the lost of important information is highly damaging. We show that the structure can actually improve the results of summarization or at least maintain state-of-the-art results, while for anchor detection it leads us to the best precision in the context of the Search and Anchoring in Video Archives task at MediaEval. The experiments were carried on written text and a more challenging corpus containing automatic transcripts of TV shows.

MOTS-CLÉS : analyse de salves lexicales, hiérarchie de fragments thématiques, résumé automatique, détection d'ancres.

KEYWORDS: burst analysis, hierarchy of topical fragments, text summarization, anchor detection.

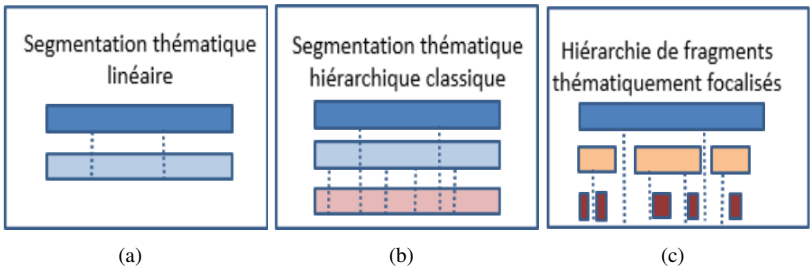


FIGURE 1 – Représentations génériques (a) d'une segmentation thématique linéaire, (b) d'une segmentation thématique hiérarchique dense classique, *versus* (c) celle d'une hiérarchie de fragments thématiquement focalisés. Les lignes verticales en pointillés illustrent les frontières des thèmes et sous-thèmes.

1 Introduction

Les algorithmes développés pour mettre au jour la structure thématique de documents ont pour objectif de détecter automatiquement les frontières qui délimitent des segments thématiquement cohérents dans un texte. Identifier une telle structure thématique est une étape essentielle pour différentes tâches du traitement automatique des langues telles que le résumé automatique, la recherche d'information, les systèmes de questions-réponses, *etc.* On peut distinguer deux formes d'organisation thématique des données : les organisations linéaire et hiérarchique. La segmentation thématique linéaire vise à structurer les données textuelles en thèmes successifs, alors que la segmentation thématique hiérarchique consiste à diviser un thème principal en sous-thèmes, qui à leur tour peuvent être divisés en sous-sous-thèmes, *etc.* Les techniques génériques de segmentation thématique exploitent traditionnellement la cohésion lexicale, ce qui leur permet d'être indépendantes du type de documents textuels traités et de ne pas nécessiter de phase d'apprentissage. Se fonder sur la cohésion lexicale signifie analyser la distribution des mots afin d'identifier les changements significatifs de vocabulaire, révélateurs de changements de thèmes.

Dans un travail récent (Simon *et al.*, 2015), nous avons étudié à quel point la segmentation thématique hiérarchique pouvait effectivement tirer parti de la cohésion lexicale et montré que cet indice était insuffisant pour retrouver des segmentations de référence. Partant de ce constat, nous avons proposé une nouvelle façon de considérer ce type d'organisation hiérarchique, passant d'une segmentation dense classique à une hiérarchie de fragments de textes thématiquement focalisés. Pour illustrer ce concept, la figure 1 présente des représentations génériques des structures thématiques possibles. La figure 1(a) correspond à une segmentation thématique linéaire, où un document est divisé en thèmes principaux. La figure 1(b) représente une segmentation thématique hiérarchique classique, dans laquelle les thèmes principaux sont divisés en sous-thèmes qui peuvent eux-mêmes être subdivisés. Enfin, la figure 1(c) illustre la nouvelle façon de concevoir la segmentation thématique hiérarchique décrite dans (Simon *et al.*, 2015). L'idée est de repérer des fragments de texte thématiquement focalisés, non nécessairement contigus (d'où cette appellation de fragment plutôt que segment), et d'organiser ces fragments à différents niveaux de façon hiérarchique.

Pour obtenir cette nouvelle structure thématique, nous avons proposé une autre façon de mesurer la cohésion lexicale, à travers une analyse de salves [1]. Une telle analyse considère les différentes

1. Nous employons dans cet article les termes « salve » ou « rafale » pour traduire les notions de *burst* et *burstiness*.

positions d'un mot dans un texte et les intervalles temporels entre ses répétitions. Elle étudie donc si un mot se situe au début, au milieu ou à la fin d'un document, et s'il apparaît fréquemment avec des occurrences successives très rapprochées ou plutôt uniformément réparties à travers le document (Sarkar *et al.*, 2005), et ce, à différents niveaux de détail. Dans (Simon *et al.*, 2015), nous avons évalué cette nouvelle façon de structurer hiérarchiquement des données textuelles grâce à une comparaison qualitative par rapport à une segmentation dense classique. Les résultats de cette évaluation montrent que la hiérarchie de fragments obtenue présente une meilleure focalisation thématique que la segmentation classique. Afin d'examiner plus avant ses capacités, nous proposons ici d'évaluer le potentiel de cette nouvelle structuration thématique – et donc plus généralement du paradigme des fragments saillants – dans les contextes applicatifs du résumé automatique par extraction et de la détection d'ancres en hyperliage vidéo (de récents résultats sur la structuration de collections montrent en effet que le langage est un élément crucial pour l'établissement de liens entre vidéos (Eskevich *et al.*, 2013)). Précisons que la détection d'ancres concerne l'identification de fragments de vidéos qu'un utilisateur peut considérer comme de bons points d'entrée pour naviguer au sein d'une collection de vidéos. Une telle évaluation applicative a pour objet de souligner le potentiel de cette nouvelle structuration à capturer – et ce, malgré la compression des données qu'elle opère en n'étant pas dense – tous les aspects importants d'un texte, qualité particulièrement fondamentale pour produire des résumés et pointer des ancres pertinentes. Pour cette double évaluation dirigée par la tâche, nous nous appuyons d'une part sur des données textuelles classiques mais également sur un jeu de données plus difficile contenant des transcriptions automatiques d'émissions de télévision.

L'article est organisé de la façon suivante : en section 2, nous présentons un survol des différentes approches de segmentation thématique hiérarchique existantes et de leurs limites, et introduisons le paradigme des fragments saillants. Nous détaillons en section 3 l'algorithme permettant de produire une hiérarchie de fragments thématiquement focalisés. Nous proposons enfin en section 4 l'évaluation par le biais d'applications de la structure produite par cet algorithme, dans le contexte du résumé automatique et de la détection d'ancres.

2 De la segmentation thématique hiérarchique classique à la production d'une hiérarchie de fragments thématiquement focalisés

Une première approche, parmi les travaux s'intéressant à la segmentation thématique hiérarchique classique, consiste à appliquer un algorithme de segmentation linéaire récursivement (Carroll, 2010; Guinaudeau, 2011). Plusieurs algorithmes de segmentation thématique linéaire ont ainsi été plongés dans un cadre hiérarchique, parmi lesquels on peut citer TextSeg (Utiyama & Isahara, 2001), LCSeg (Galley *et al.*, 2003), BayesSeg (Eisenstein & Barzilay, 2008), C99 (Choi, 2000) ou encore CWM (Choi *et al.*, 2001). Seules quelques études se sont penchées sur la modélisation explicite de la structure thématique hiérarchique. Parmi elles, (Eisenstein, 2009) décrit HierBayes, algorithme non supervisé formalisé dans un cadre probabiliste bayésien. Son principe sous-jacent est que chaque mot d'un texte est représenté par un modèle de langage estimé sur une portion plus ou moins importante de texte. Une pyramide de modèles de langage est alors construite, les modèles de haut niveau expliquant les mots dans de grandes parties du document, tandis que les modèles de langue de bas niveau expliquent un ensemble local de mots. Pour produire la segmentation hiérarchique, l'algorithme cherche à maximiser la cohésion lexicale des segments à chaque niveau de la hiérarchie, imposant

que les frontières à un niveau supérieur dans cette hiérarchie soient alignées avec celles calculées aux niveaux inférieurs. Le point faible de cette approche est qu'elle ne peut pas traiter les cas de segments thématiques de longueurs variables, et qu'elle nécessite des informations préalables sur la longueur attendue des segments à chaque niveau et sur le nombre de niveaux de la hiérarchie. Les auteurs de (Kazantseva & Szpakowicz, 2014) proposent quant à eux d'utiliser le modèle graphique de propagation d'affinités hiérarchique présenté dans (Givoni *et al.*, 2011) pour extraire la structure thématique hiérarchique. Cette approche requiert également de l'information *a priori* sur la granularité de la segmentation. Dans (Moens & Busser, 2001), les auteurs décrivent une approche ascendante qui consiste à connecter les segments qui ont un lien hiérarchique pour inférer la structure d'un texte. Ils s'appuient sur des heuristiques génériques portant sur les chaînes lexicales pour extraire le thème principal de chaque phrase, telles que la position dans la phrase, la persistance par rapport aux phrases précédentes, les références pronominales... Ces heuristiques sont considérées comme applicables à plusieurs langues à ordre sujet-verbe-objet, tels que l'anglais, le français ou le néerlandais. Les débuts, interruptions et fins de chaînes lexicales sont vus comme porteurs d'informations précieuses sur les frontières des thèmes.

Les techniques décrites précédemment ont été principalement appliquées à des textes standards et présentent toujours des limites, non abordées dans la littérature. Dans le cas de l'application récursive d'un algorithme de segmentation linéaire, il est ainsi difficile de décider quand arrêter de segmenter. De plus, les erreurs de segmentation à un niveau sont propagées aux autres niveaux, ce qui rend l'évaluation du potentiel réel de la méthode difficile. Par ailleurs, le nombre de mots considérés – de plus en plus faible au fur et à mesure de l'application récursive de l'algorithme – et l'écart thématique de moins en moins marqué entre les sous-segments à distinguer – des sous-sous-thèmes d'un thème initial étant moins différents entre eux que ce thème des autres thèmes initialement repérés – questionnent fortement les capacités des approches fondées sur la répétition de vocabulaire. Un autre aspect important généralement non pris en compte lors de l'application récursive d'algorithmes de segmentation linéaire est que les mots qui ont contribué à la segmentation d'un niveau dans la hiérarchie devraient avoir une importance différente pour l'obtention des segments à d'autres niveaux. Par ailleurs, certains des algorithmes qui extraient directement la hiérarchie ont besoin d'informations sur le niveau de granularité et les longueurs attendues des segments, informations non disponibles dans un scénario réel.

Toutes ces limites nous ont conduits à envisager d'une autre façon la segmentation thématique hiérarchique et à proposer, dans (Simon *et al.*, 2015), une nouvelle structuration hiérarchique. Cette structuration est obtenue à l'aide de l'algorithme HTFF (pour *hierarchy of topically focused fragments*) qui, plutôt que de chercher les changements de thèmes pour aboutir à une segmentation dense classique, extrait des fragments thématiquement focalisés organisés de façon hiérarchique, en exploitant le phénomène de salves d'occurrences. Il repose sur le fait que la présence de salves lexicales indique un focus thématique marqué. Dans (Altmann *et al.*, 2009), les auteurs soulignent que les mots en rafales caractérisent mieux les thèmes du discours. De même, Kleinberg, dans (Kleinberg, 2002), soutient cette idée et suggère qu'un thème devrait être caractérisé par une « explosion d'activité ». Se fonder sur le phénomène de salves des mots signifie aussi aller au-delà du modèle sac-de-mots qui est un descripteur pauvre des occurrences des mots (Sarkar *et al.*, 2005; Madsen *et al.*, 2005; Lijffijt *et al.*, 2011). D'ailleurs, des travaux sur les lois statistiques en langue ont présenté des modèles de détection de rafales qui analysent les motifs distributionnels des mots (Sarkar *et al.*, 2005; Madsen *et al.*, 2005) pour surmonter les déficiences du modèle multinomial. Si l'analyse de salves a été soulignée comme potentiellement intéressante et si des modèles de détection des rafales ont été proposés, ces derniers n'ont en revanche pas été exploités dans un cadre de structuration thématique des textes, ce que nous

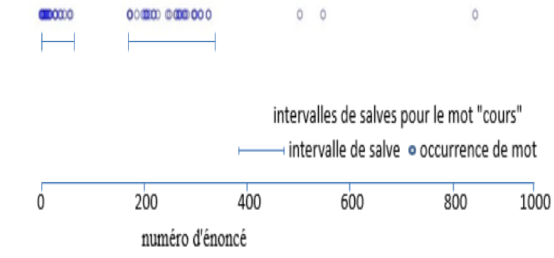
avons fait dans (Simon *et al.*, 2015) avec l'algorithme HTFF que nous décrivons plus en détail dans la section suivante.

3 L'algorithme HTFF

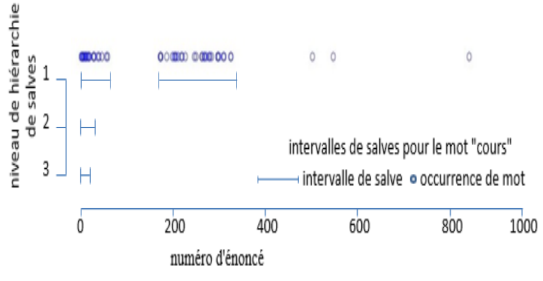
L'algorithme HTFF prend en compte l'information positionnelle des mots, ce qui signifie qu'il s'intéresse au fait qu'un mot apparaisse avec des occurrences successives proches ou, au contraire, avec des intervalles longs entre ses répétitions. Chaque fois qu'un mot arrive par salve, on considère que quelque chose de notable et important se passe au niveau textuel par rapport au reste du texte. Cet aspect notable est représenté par un regain d'activité dans l'utilisation d'un mot qui va être ensuite associé à un thème. L'algorithme HTFF capture les rafales d'activité en s'appuyant sur l'algorithme de Kleinberg (Kleinberg, 2002) qui, pour chaque mot, reconnait s'il présente des salves et produit, dans ce cas, une représentation imbriquée de l'ensemble de ses salves (c-à-d une hiérarchie d'intervalles de salves). Les mots en rafales tendent à être caractérisés par des intervalles de temps longs entre leurs apparitions suivis par des intervalles de temps courts, tandis que les autres mots présentent des intervalles d'apparition avec moins de différences (Lijffijt *et al.*, 2011). La figure 2 illustre la sortie de l'algorithme de Kleinberg, algorithme qui fonctionne sans connaissance *a priori* du nombre de niveaux à produire et détermine directement les frontières des intervalles de salves. En 2(a), deux intervalles de salves détectés pour le mot « cours » dans une transcription automatique d'une émission TV Envoyé Spécial sont illustrés. Les cercles représentent les occurrences du mot dans l'ensemble de la transcription ; ces occurrences conduisent à la formation de deux intervalles de salves, marquant le début et la fin des groupes de souffle de la transcription dans lesquels ce mot apparaît de façon notable. Un aspect particulièrement intéressant de l'algorithme de Kleinberg est qu'il produit en fait une hiérarchie d'intervalles de salves pour chaque mot. Comme cela est représenté en 2(b), le mot « cours » présente, au début du premier intervalle de salve du niveau 1, une fréquence de répétition localement plus élevée. Modéliser les différent intervalles temporels entre occurrences conduit à l'obtention de la représentation imbriquée des intervalles de salves à différents niveaux. Chaque niveau de cette représentation correspond à une certaine intensité dans l'utilisation du mot, le niveau le plus élevé indiquant une intensité plus forte avec des écarts plus petits entre les répétitions du mot.

En exploitant l'algorithme de Kleinberg fournissant une hiérarchie des intervalles de salves pour chaque mot d'une collection de documents, nous avons proposé (Simon *et al.*, 2015) l'algorithme HTFF qui réalise un regroupement agglomératif des intervalles de salves pour construire une organisation thématique de chaque document. La figure 3 représente les étapes de cet algorithme permettant de créer la hiérarchie de fragments thématiquement focalisés, sur un exemple à deux mots *A* et *B*. Tout d'abord, les intervalles de salves des différents mots à un niveau de la hiérarchie sont considérés et, chaque fois que des intervalles se chevauchent, ils sont regroupés pour former un intervalle plus grand. Chaque nouvel intervalle peut être représenté par les mots en rafales qui ont conduit à sa formation. Ce regroupement peut être fait en parallèle pour chaque niveau dans la hiérarchie des salves, ce qui conduit à la création de nouveaux intervalles à différents niveaux de la hiérarchie. Les intervalles en bleu (c-à-d les résultats du regroupement) vont correspondre aux fragments focalisés de l'algorithme HTFF.

Nous donnons en figure 4 un exemple de la structure de sortie obtenue en appliquant HTFF sur une transcription automatique de l'émission TV de la BBC *Castle in the country*, avec des images-clés issues des fragments formés au niveau le plus haut de la hiérarchie. On peut remarquer un fragment



(a)



(b)

FIGURE 2 – Sortie de l'algorithme de Kleinberg : (a) les intervalles de salves du mot « cours » ; (b) la représentation imbriquée de tous les intervalles de salves détectés pour ce mot.

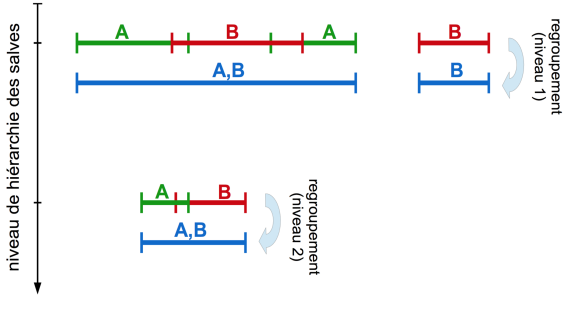


FIGURE 3 – Obtention de la hiérarchie de fragments thématiquement focalisés par regroupement des fragments qui se chevauchent à chaque niveau de la hiérarchie des salves.

thématiquement focalisé de 0.01 à 10.5 (format du temps : minutes.secondes) qui peut être caractérisé par les mots dont le regroupement des intervalles de salves forme ce fragment (par exemple *ballroom*, *king*...). Si on zoome au sein de ce fragment, d'autres fragments thématiquement focalisés plus petits apparaissent, tels celui de 0.01 à 1.50 caractérisé par certains des mots ayant contribué à la création du niveau inférieur (c-à-d celui sur lequel on a zoomé). On peut également observer l'absence de contiguïté des fragments (par exemple, de 1.50 à 2.29). Notre approche réalise en fait une compression des données initiales, ne conservant que l'information saillante. L'information laissée

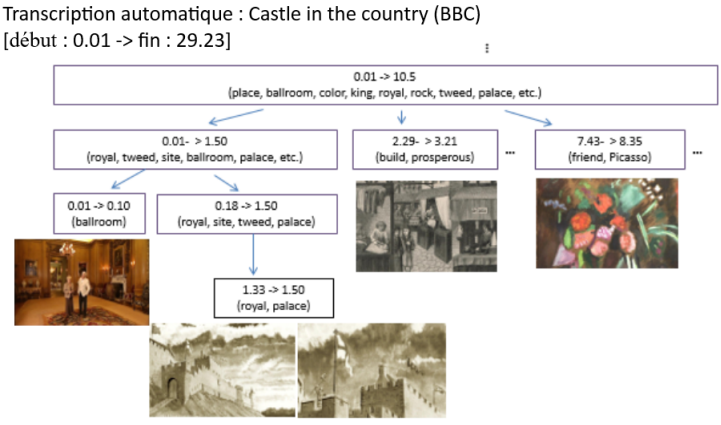


FIGURE 4 – Exemple de sortie de HTFF appliqué à la transcription automatique de l'émission TV *Castle in the country*.

de côté est considérée comme du simple remplissage, c-à-d des portions de données qui n'apportent pas d'information additionnelle intéressante. Dans la section suivante, nous nous penchons sur l'évaluation de cet algorithme et de sa production dans le contexte de deux cadres applicatifs : le résumé automatique et la détection d'ancres.

4 Évaluation par le biais d'applications

Jusqu'à présent une preuve de concept pour l'algorithme HTFF a été réalisée grâce à une comparaison qualitative et quantitative avec une segmentation dense traditionnelle, pour laquelle une segmentation hiérarchique de référence existe. L'évaluation directe de cet algorithme est en effet quelque chose de difficile car il n'existe pas de vérité-terrain correspondant à la structure hiérarchique produite, et la production d'un guide d'annotation pour la créer est elle-même un défi. Par conséquent, nous proposons ici d'étudier l'impact de la structure obtenue dans le contexte du résumé automatique d'une part (sous-section 4.1) et de la détection d'ancres d'autre part (sous-section 4.2). Dans cette nouvelle structure hiérarchique des données non dense, c'est-à-dire avec des fragments non contigus, une partie du contenu à divers niveaux de détail est éliminée. Pour les tâches au sein desquelles nous voulons utiliser cette nouvelle structure, il est important cependant que l'information principale soit conservée. Par conséquent, ces évaluations par le biais d'applications ont pour objet de montrer qu'en utilisant HTFF ce qui est important dans les données est bien préservé.

Les corpus utilisés pour le résumé automatique sont d'une part les transcriptions de 7 émissions de TV Envoyé Spécial (automatiques pour les 7 et manuelles pour 4 d'entre elles) et, d'autre part, un corpus de romans [2] proposé pour cette tâche dans (Kazantseva, 2006). Ce dernier corpus contient 20 chapitres issus de plusieurs romans du 19^e siècle et début du 20^e, et a été scindé en deux groupes de 10 chapitres. Les chapitres de chaque groupe ont été annotés manuellement au niveau phrase selon un ensemble d'instructions [3] par trois personnes, plus un annotateur commun pour les deux groupes

2. Ce corpus est disponible à http://www.eecs.uottawa.ca/∼ankazant .
3. Le guide d'annotation est disponible à http://www.site.uottawa.ca/∼ankazant/instructions.zip .

(l'auteur de (Kazantseva, 2006)). Quatre résumés manuels sont donc produits pour chaque chapitre de chacun des deux groupes. Le corpus de transcription d'émissions TV est plus difficile car les transcriptions ne respectent pas les normes de l'écrit : pas de signes de ponctuation, pas de majuscules, une structuration non pas en phrases mais en groupes de souffle, et des erreurs de transcription. Nous avons choisi ce corpus pour montrer que l'algorithme HTFF est robuste aux données bruitées.

Pour la tâche de détection d'ancres, l'évaluation correspond à la nouvelle tâche *Search and Anchoring in Video Archives* (SAVA) de la campagne d'évaluation MediaEval (Eskevich *et al.*, 2015). Dans cette tâche, les participants doivent définir les segments-ancres pour 33 vidéos d'émissions TV de la BBC. Les émissions ont une durée moyenne de 45 minutes. Les vidéos ont été enregistrées entre le 1^{er} avril 2008 et le 31 juillet 2008 et sont très variées : des actualités, des séries TV, des documentaires, des émissions pour enfants, du sport, des divertissements... En plus du contenu des vidéos, les organisateurs ont également fourni les transcriptions manuelles réalisées par des experts, ainsi que les transcriptions automatiques obtenues à l'aide de différents systèmes de reconnaissance automatique de la parole, parmi lesquels les systèmes du LIMSI (Gauvain *et al.*, 2002) et du LIUM (Rousseau *et al.*, 2014). Pour nos expérimentations, nous sommes appuyés sur les transcriptions manuelles et celles du système LIMSI.

Toutes les données ont été lemmatisées avec TreeTagger, et seuls les noms, les verbes non modaux et les adjectifs ont été conservés.

4.1 Résumé automatique

Étant donnée la nouvelle structure hiérarchique produite par HTFF, nous pouvons nous confronter à la tâche de résumé automatique selon deux angles. D'une part, nous pouvons évaluer si un générateur état-de-l'art de résumés automatiques peut bénéficier ou pas de la compression des textes produite par HTFF. D'autre part, nous pouvons produire directement des résumés, en exploitant HTFF pour en créer à différents niveaux de détail. Cette approche nécessite une vérité-terrain de résumés eux aussi à différents niveaux de détail pour pouvoir évaluer la qualité de la hiérarchie dans son ensemble. Ne disposant pas d'une telle vérité-terrain, nous pourrions nous limiter à un seul niveau de la hiérarchie. Cependant, un résumé de taille limitée doit être proposé afin d'être en mesure de le comparer à celui produit par une autre méthode. Ceci est en effet nécessaire car les mesures utilisées pour l'évaluation de résumés pourraient être biaisées en faveur de résumés plus longs. Face à cette difficulté, nous nous concentrons ici sur le seul premier angle, une manière appropriée d'utiliser directement la hiérarchie pour la production de résumés restant encore à trouver.

Pour montrer si l'information importante est effectivement conservée dans la structure issue de HTFF et tester si un système produisant des résumés peut tirer profit de cette compression du texte, notre expérience a consisté à utiliser le générateur proposé dans (Gillick & Favre, 2009) (nommé par la suite ILP-sum [4]) à la fois sur les textes originaux du corpus de romans et sur les textes compressés (c-à-d le niveau 1 de la hiérarchie de fragments thématiquement focalisés, qui correspond ici à l'élimination de 45% des données en nombre de mots). La qualité des résumés est évaluée classiquement en utilisant les métriques Rouge (Lin, 2004). Les scores de rappel, précision et F1-mesure obtenus pour Rouge-1, Rouge-2, Rouge-3, Rouge-4, Rouge-L et Rouge-W sont fournis dans le tableau 1. Rouge-n compare les n-grammes contenus dans le résumé produit à ceux des résumés de la vérité-terrain. Rouge-L considère la plus longue sous-séquence commune (LCS) entre le résumé généré et la vérité-terrain,

4. https://github.com/boudinfl/sume (Boudin *et al.*, 2015).

							Racineur					
	Original			Compressé			Original			Compressé		
Mesure	R	P	F	R	P	F	R	P	F	R	P	F
Rouge-1	0.451	0.442	0.446	0.456	0.448	0.452	0.467	0.458	0.463	0.472	0.464	0.468
Rouge-2	0.175	0.171	0.173	0.188	0.184	0.186	0.177	0.174	0.176	0.19	0.187	0.188
Rouge-3	0.127	0.124	0.125	0.144	0.142	0.143	0.127	0.125	0.126	0.145	0.142	0.143
Rouge-4	0.115	0.112	0.113	0.134	0.131	0.132	0.115	0.112	0.113	0.134	0.131	0.133
Rouge-L	0.429	0.421	0.425	0.434	0.426	0.43	0.443	0.434	0.438	0.447	0.439	0.443
Rouge-W	0.128	0.239	0.167	0.19	0.191	0.191	0.131	0.245	0.171	0.135	0.253	0.176

TABLE 1 – Rappel, précision et F1-mesure pour Rouge 1-4, Rouge-L et Rouge-W, mesures obtenues avec ILP-sum sur le corpus de romans, versions originale et compressée. Les résultats pour l'entrée racinisée sont également fournis.

tandis que Rouge-W donne aux appariements consécutifs de longueur L dans une LCS un poids de L^{poids} au lieu de L. Pour notre évaluation, le poids habituel de 1,2 a été retenu, et le nombre de mots d'un résumé produit est limité à la taille moyenne des résumés de référence. *Compressé* dans le tableau fait référence aux résumés obtenus sur les données compressées en utilisant la hiérarchie de fragments thématiquement focalisés. Tous les résultats sur les données compressées sont plus élevés que ceux obtenus quand les données entières sont considérées. Les différences ne sont cependant pas statistiquement significatives mais l'absence de dégradation des résultats montre que l'on conserve bien l'information importante. Les valeurs plus élevées peuvent s'expliquer par le fait que la nouvelle structure aide à amener à la surface de l'information qui ne serait pas incluse si le texte complet était pris en compte. Avec la compression, une partie des données où les mots importants ont tendance à disparaître est ignorée, tandis que lorsque l'on considère l'ensemble du texte, ces données sont toujours perçues comme informatives. On est donc à même de se concentrer sur ce qui est important et à ne pas prendre en compte le bruit.

Dans (Louis & Nenkova, 2013), les auteurs proposent plusieurs stratégies d'évaluation du contenu de résumés automatiques en l'absence de référence, ce qui est notre cas pour les émissions TV. Elles montrent qu'en quantifiant la similarité entre le texte source et son résumé avec des mesures choisies de façon appropriée, elles sont à même de reproduire correctement des évaluations humaines. Nous avons donc considéré les mesures qu'elles proposent pour évaluer la qualité des résumés produits pour notre second corpus. Comme précédemment, nous avons généré des résumés à l'aide d'ILP-sum à la fois sur les données originales et sur leur version compressée grâce à HTFF (la compression est ici de 7% au niveau 1 de la hiérarchie sur les transcriptions automatiques, taux moindre que pour les romans s'expliquant par le fait que les émissions contiennent plusieurs reportages sur des sujets divers, identifiés à travers des intervalles de salves étendus au premier niveau, la compression atteignant 42% au deuxième niveau). La qualité des résumés est évaluée en se fondant sur la distribution des mots dans l'entrée originale et les résumés, les bons résumés par extraction ayant tendance à avoir un contenu similaire à celui de l'entrée (Louis & Nenkova, 2013). Parmi les mesures proposées dans l'article cité, la divergence de Jensen Shannon (JSd) obtient les meilleures corrélations entre les scores automatiques et manuels. L'intuition derrière cette mesure est que la distance entre deux distributions ne peut pas être très différente de la moyenne des distances par rapport à leur distribution moyenne. JSd est définie par :

$$JSd(P||Q) = \frac{1}{2}[D(P||A) + D[Q||A]] \ ,$$

où P et Q sont les distributions, dans notre cas, des mots dans les entrée et résumé, et $A = \frac{P+Q}{2}$ est la distribution moyenne de P et Q. La divergence (de Kullback Leibler) entre deux distributions de

	Émissions TV (automatique)		Émissions TV (manuel)		Romans	
Système	JSd moy	JSd moy lissée	JSd moy	JSd moy lissée	JSd moy	JSd moy lissée
Entrée–Original	0.54	0.46	0.53	0.45	0.36	0.32
Entrée–Compressé	0.53	0.46	0.53	0.46	0.37	0.33
Entrée–Vérité-terrain	–	–	–	–	0.4	0.35
Vérité-terrain–Original	–	–	–	–	0.47	0.45
Vérité-terrain–Compressé	–	–	–	–	0.46	0.44

TABLE 2 – Scores JSd obtenus sur les corpus d'émissions de TV et de romans.

probabilités P et A est donnée par :

$$D(P||A) = \sum_w p_P(w) log_2 \frac{p_P(w)}{p_A(w)} \quad .$$

Dans le tableau 2, nous présentons les résultats obtenus pour JSd pour les résumés produits à partir des émissions de TV et des romans. Nous avons utilisé l'outil fourni par les auteurs de (Louis & Nenkova, 2013) [5]. Nous donnons à la fois les valeurs pour la version lissée ou pas de JSd. *Entrée* fait référence au texte initial à résumer ; *Original* et *Compressé* sont les résumés obtenus en appliquant ILP-sum sur le texte entier et sur sa version compressée respectivement. Pour le corpus de romans, nous donnons aussi les scores de JSd impliquant la vérité-terrain (c-à-d les résumés manuels). Enfin *manuel* et *automatique* concernent respectivement les transcriptions manuelles et automatiques des (respectivement 4 et 7) émissions TV. Les valeurs de divergence les plus basses sont les meilleures. Comme on peut le voir, les différences entre les méthodes sont faibles ou inexistantes en termes de mesure JSd. Les évaluations proposées jusqu'à présent montrent donc que nous réussissons à conserver, dans la nouvelle structure de fragments thématiquement focalisés, l'information importante contenue dans les données initiales.

4.2 Détection d'ancres

La détection d'ancres, comme évoqué en introduction, a pour objet l'extraction automatique de segments de vidéos pouvant servir de points de départ à l'exploration d'une collection de vidéos, des utilisateurs étant intéressés par l'acquisition d'informations supplémentaires à partir de ces ancres. Notre approche pour détecter ces ancres consiste à structurer chaque vidéo en une hiérarchie de fragments thématiquement focalisés grâce à l'algorithme HTFF. HTFF s'appliquant à des données textuelles, nous exploitons par conséquent ici les transcriptions automatiques de la parole contenue dans les vidéos, ainsi que les sous-titres manuels fournis (Gauvain *et al.*, 2002). La structure obtenue permet d'extraire des fragments avec des points d'entrée précis et à différents niveaux de détail. Une fois extraits, une sélection parmi eux peut être faite. L'avantage de l'utilisation de HTFF est que l'algorithme aide à identifier l'information saillante des vidéos, ignorant donc celle non pertinente. De plus, grâce à la représentation hiérarchique, les fragments fournis en résultats peuvent être de granularités différentes, c-à-d plus spécifiques ou plus généraux. Des ancres couvrant un thème général ou divers points de vue sur un même sujet peuvent ainsi être proposées.

Après obtention des fragments thématiquement focalisés, pour chaque vidéo de laquelle des ancres doivent être extraites, nous réalisons une analyse de leur contenu afin de proposer les meilleures

5. http://homepages.inf.ed.ac.uk/alouis/IEval2.html .

ancres par vidéo. Pour ce faire, nous utilisons la mesure de cohésion probabiliste $C(S_i)$, à la base de l'algorithme de segmentation thématique (Utiyama & Isahara, 2001), permettant d'ordonner les fragments de la hiérarchie, mesure définie par

$$C(S_i) = log \prod_{j=1}^{n_i} \frac{f_i(w_j^i) + 1}{n_i + k} \ ,$$

où n_i est le nombre d'occurrences de mots dans le fragment S_i, $f_i(w_j^i)$ est le nombre d'occurrences du mot w_j^i dans S_i, et k est la taille du vocabulaire $\mathcal{V}$. Cette mesure favorise les fragments petits et homogènes, augmentant quand les mots sont répétés et diminuant en conséquence quand ils sont distincts. Ce choix de privilégier les fragments les plus cohérents repose sur l'idée qu'un fragment focalisé sur un thème précis sera plus pertinent pour remplir la fonction d'ancre qu'un segment mélangeant différents thèmes. Nous ordonnons donc les fragments de tous les niveaux de la hiérarchie grâce à la mesure de cohésion. Les fragments de plus de 2 minutes sont éliminés afin de favoriser ceux plus précis.

Utiliser HTFF pour détecter des ancres n'assure pas qu'un nombre minimum d'ancres soit détecté pour une vidéo. Plus ou moins d'ancres sont proposées selon les vidéos, ce qui est réaliste puisque le nombre d'ancres détectables dans une vidéo dépend de l'information saillante qu'elle contient. Le nombre moyen d'ancres par vidéo (quel que soit le niveau de la hiérarchie dont ils proviennent) obtenu en exploitant les sous-titres est de 18,9 avec un intervalle de confiance à 95% de [18,56 ; 19,25]. Quand les transcriptions automatiques sont utilisées, ce nombre passe à 19,03 avec un intervalle de confiance à 95% de [18,4 ; 19,65]. Nous nous intéressons ici à un système visant une forte précision plus qu'un rappel élevé. Nous avons choisi de ne pas proposer plus de 20 ancres par vidéo.

Les résultats obtenus (précision à 10, rappel et MRR (*mean reciprocal rank*)) par tous les participants de la tâche SAVA sont rassemblés dans le tableau 3. La méthodologie d'évaluation et les systèmes mentionnés ci-après sont décrits dans (Larson *et al.*, 2015). L'évaluation par *crowd-sourcing* a concerné les 25 ancres de rangs les plus élevés de chaque soumission, le reste des ancres d'un système ne recevant un jugement qu'en cas de classement parmi les 25 premiers d'autres soumissions. Nos résultats correspondent au système *A*, et nous avons été les seuls à exploiter les transcriptions automatiques. On peut observer que les systèmes *A* et *B* obtiennent les meilleurs scores de précision, notre système atteignant ce meilleur score sur les transcriptions automatiques alors que le *B* l'obtient sur les sous-titres. Notre approche présente en revanche de moins bons résultats sur les sous-titres manuels ; nous pensons que cela est dû au fait que les ancres sélectionnées dans ce cas sont plus courtes. La durée moyenne est alors de 11,46 secondes, avec un intervalle de confiance à 95% de [11,5 ; 11,86], alors que pour les transcriptions, la durée moyenne des ancres est de 22,92 secondes (intervalle de confiance à 95% de [21,48 ; 24,35]). Par ailleurs, un nombre plus réduit d'ancres est proposé pour les sous-titres, ce que nous croyons lié au fait qu'il y ait plus de répétitions de mots dans les sous-titres que dans les transcriptions, qui contiennent d'ailleurs des mots erronés. Pour ce qui est du rappel, le système *B* a les meilleures performances. Le mode d'évaluation choisi explique au moins partiellement le résultat plus faible de notre système, ne retournant qu'un maximum de 20 ancres par vidéo, sur ce point ; le rappel est en effet calculé comme la proportion d'ancres, parmi les ancres jugées pertinentes au sein de celles soumises par l'ensemble des systèmes, qu'un système donné a proposé, ce qui peut prêter à discussions. Enfin, le MRR est également plus faible pour notre système que pour les autres participants, ce qui signifie que leur première ancre pertinente apparaît plus haut dans leurs classements que dans notre cas. Ceci ne remet pas en cause la précision de nos résultats mais l'utilisation de la seule cohésion lexicale comme critère d'ordonnancement des ancres.

Système	Type de données	Précision@10	Rappel	MRR
A	Sous-titres manuels	0.469	0.38	0.73
A	Transcriptions LIMSI	**0.557**	0.435	0.77
B	Sous-titres manuels	**0.557**	**0.474**	**0.87**
C	Sous-titres manuels	0.31	0.27	0.83

TABLE 3 – Résultats en précision, rappel et MRR pour tous les systèmes de la tâche SAVA.

5 Conclusion

Dans cet article, nous avons étudié, à travers son intérêt pour le résumé automatique et la détection d'ancres dans des vidéos, le potentiel d'une nouvelle structure thématique non dense extraite de données textuelles, composée d'une hiérarchie de fragments thématiquement focalisés. L'évaluation guidée par la tâche dans le cas du résumé automatique a montré que l'utilisation de cette représentation qui effectue une compression des données n'a pas d'incidence négative sur le résultat produit, et à même tendance à l'améliorer. Ceci valide l'intérêt du paradigme des fragments saillants et justifie la poursuite de travaux sur la conception d'un algorithme de production de résumés directement à partir de la hiérarchie. Outre l'évaluation sur de plus grands ensembles de données (par exemple, les données de DUC et TAC), une autre perspective concerne la recherche d'une manière adéquate pour évaluer les résumés qui pourraient être produits en s'appuyant sur chacun des niveaux de la hiérarchie. La nouvelle structure a également démontré son intérêt dans le contexte de la détection d'ancres. Nous avons obtenu grâce à elle des résultats aussi élevés en précision sur les transcriptions automatiques que ceux atteints par le meilleur système sur les sous-titres manuels. Cependant, lorsque nous appliquons notre approche sur ces sous-titres, la précision diminue, ce que nous pensons dû d'une part au nombre plus restreint d'ancres proposées et, d'autre part, à la durée plus limitée de celles-ci. Une façon de pallier ce problème consisterait à améliorer l'évaluation de la pertinence des ancres et à proposer des segments plus longs comme meilleures ancres. Par ailleurs, pour ces deux tâches applicatives, une méthodologie d'exploitation des relations hiérarchiques entre fragments reste encore à déterminer.

Nous avons aussi pour objectif d'améliorer notre algorithme HTFF. Le modèle de détection des salves pourrait gagner à prendre en compte des relations sémantiques ou d'autres modèles plus complexes d'analyse de distributions d'un mot. En effet, depuis la proposition de l'algorithme par Kleinberg en 2002, d'autres modèles ont vu le jour (Pui *et al.*, 2005; Sarkar *et al.*, 2005; Madsen *et al.*, 2005; Altmann *et al.*, 2009). Déterminer une façon de les inclure au sein de l'algorithme de Kleinberg serait intéressant pour conserver la caractéristique de structuration hiérarchique des salves de celui-ci.

Références

ALTMANN E. G., PIERREHUMBERT J. B. & MOTTER A. E. (2009). Beyond word frequency : Bursts, lulls, and scaling in the temporal distributions of words. *PLoS ONE*, **4**(11).

BOUDIN F., MOUGARD H. & FAVRE B. (2015). Concept-based summarization using integer linear programming : From concept pruning to multiple optimal solutions. In *Empirical Methods in Natural Language Processing*, p. 1914–1918.

CARROLL L. (2010). Evaluating hierarchical discourse segmentation. In *11th International Conference of the North American Chapter of the Association for Computational Linguistics*, p. 993–1001.

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

CHOI F. Y. Y., WIEMER-HASTINGS P. & MOORE J. (2001). Latent semantic analysis for text segmentation. In *Empirical Methods in Natural Language Processing*, p. 109–117.

EISENSTEIN J. (2009). Hierarchical text segmentation from multi-scale lexical cohesion. In *10th International Conference of the North American Chapter of the Association for Computational Linguistics*, p. 353–361.

EISENSTEIN J. & BARZILAY R. (2008). Bayesian unsupervised topic segmentation. In *Empirical Methods in Natural Language Processing*, p. 334–343.

ESKEVICH M., ALY R., ORDELMAN R., RACCA D. N., CHEN S. & JONES G. J. F. (2015). SAVA at MediaEval 2015 : Search and Anchoring in Video Archives. In *MediaEval 2015 Workshop*.

ESKEVICH M., JONES G. J., ALY R., ORDELMAN R. J., CHEN S., NADEEM D., GUINAUDEAU C., GRAVIER G., SÉBILLOT P., DE NIES T., DEBEVERE P., VAN DEWALLE R., GALUSCAKOVA P., PECINA P. & LARSON M. (2013). Multimedia information seeking through search and hyperlinking. In *3rd International Conference on Multimedia Retrieval*, p. 287–294.

GALLEY M., MCKEOWN K., FOSLER-LUSSIER E. & JING H. (2003). Discourse segmentation of multi-party conversation. In *41st Annual Meeting of the Association for Computational Linguistics*, p. 562–569.

GAUVAIN J.-L., LAMEL L. & ADDA G. (2002). The LIMSI broadcast news transcription system. *Speech Communication*, **37**(1–2), 89–108.

GILLICK D. & FAVRE B. (2009). A scalable global model for summarization. In *Workshop on Integer Linear Programming for Natural Language Processing*, p. 10–18.

GIVONI I. E., CHUNG C. & FREY B. J. (2011). Hierarchical affinity propagation. In *27th Conference on Uncertainty in Artificial Intelligence*, p. 238–246.

GUINAUDEAU C. (2011). *Structuration automatique de flux télévisuels*. PhD thesis, INSA Rennes, France.

KAZANTSEVA A. (2006). Automatic summarization of short fiction. Master's thesis, University of Ottawa, Canada.

KAZANTSEVA A. & SZPAKOWICZ S. (2014). Hierarchical topical segmentation with affinity propagation. In *25th International Conference on Computational Linguistics*, p. 37–47.

KLEINBERG J. (2002). Bursty and hierarchical structure in streams. In *8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, p. 91–101.

M. A. LARSON, B. IONESCU, M. SJÖBERG, X. ANGUERA, J. POIGNANT, M. RIEGLER, M. ESKEVICH, C. HAUFF, R. F. E. SUTCLIFFE, G. J. F. JONES, Y. YANG, M. SOLEYMANI & S. PAPADOPOULOS, Eds. (2015). *Working Notes Proceedings of the MediaEval 2015 Workshop, Wurzen, Germany, September 14-15, 2015*, volume 1436 of *CEUR Workshop Proceedings*. CEUR-WS.org.

LIJFFIJT J., PAPAPETROU P., PUOLAMÄKI K. & MANNILA H. (2011). Analyzing word frequencies in large text corpora using inter-arrival times and bootstrapping. In *European Conference on Machine Learning and Knowledge Discovery in Databases*, p. 341–357.

LIN C.-Y. (2004). Rouge : A package for automatic evaluation of summaries. In *Text Summarization Branches Out, ACL Workshop*, p. 74–83.

LOUIS A. & NENKOVA A. (2013). Automatically assessing machine summary content without a gold-standard. *Computational Linguistics*, **39**(2), 267–300.

MADSEN R. E., KAUCHAK D. & ELKAN C. (2005). Modeling word burstiness using the Dirichlet distribution. In *22nd International Conference on Machine Learning*, p. 545–552.

MOENS M.-F. & BUSSER R. D. (2001). Generic topic segmentation of document texts. In *24th International Conference on Research and Developement in Information Retrieval*, p. 418–419.

PUI G., FUNG C., XU J., PHILIP Y., YU S. & LU H. (2005). Parameter free bursty events detection in text streams. In *31st International Conference on Very Large Data Bases*, p. 181–192.

ROUSSEAU A., DELÉGLISE P. & ESTÈVE Y. (2014). Enhancing the TED-LIUM corpus with selected data for language modeling and more TED talks. In *9th International Conference on Language Resources and Evaluation*, p. 3935–3939.

SARKAR A., GARTHWAITE P. H. & DE ROECK A. (2005). A Bayesian mixture model for term re-occurrence and burstiness. In *9th Conference on Computational Natural Language Learning*, p. 48–55.

SIMON A., SÉBILLOT P. & GRAVIER G. (2015). Hierarchical topic structuring : From dense topic segmentation to topically focused fragments via burst analysis. In *10th Recent Advances in Natural Language Processing Conference*, p. 588–595.

UTIYAMA M. & ISAHARA H. (2001). A statistical model for domain-independent text segmentation. In *39th Annual Meeting on the Association for Computational Linguistics*, p. 499–506.