TALN 2008, Avignon, 9–13 juin 2008

Un modèle multi-sources pour la segmentation en sujets de journaux radiophoniques

Stéphane Huet[1], Guillaume Gravier[2], Pascale Sébillot[3]
Institut de Recherche en Informatique et Systèmes Aléatoires, Rennes
(1) Université de Rennes 1, (2) CNRS, (3) INSA de Rennes
{stephane.huet,guillaume.gravier,pascale.sebillot}@irisa.fr

Résumé. Nous présentons une méthode de segmentation de journaux radiophoniques en sujets, basée sur la prise en compte d'indices lexicaux, syntaxiques et acoustiques. Partant d'un modèle statistique existant de segmentation thématique, exploitant la notion de cohésion lexicale, nous étendons le formalisme pour y inclure des informations d'ordre syntaxique et acoustique. Les résultats expérimentaux montrent que le seul modèle de cohésion lexicale ne suffit pas pour le type de documents étudié en raison de la taille variable des segments et de l'absence d'un lien direct entre segment et thème. L'utilisation d'informations syntaxiques et acoustiques permet une amélioration substantielle de la segmentation obtenue.

Abstract. We present a method for story segmentation of radio broadcast news, based on lexical, syntactic and audio cues. Starting from an existing statistical topic segmentation model which exploits the notion of lexical cohesion, we extend the formalism to include syntactic and acoustic knwoledge sources. Experimental results show that the sole use of lexical cohesion is not efficient for the type of documents under study because of the variable size of the segments and the lack of direct relation between topics and stories. The use of syntactics and acoustics enables a consequent improvement of the quality of the segmentation.

Mots-clés : segmentation en sujets, corpus oraux, cohésion lexicale, indices acoustiques, indices syntaxiques.

Keywords: story segmentation, spoken documents, lexical cohesion, acoustic cues, syntactic cues.

1 Introduction

Le traitement de flux de données sonores ou encore multimédias à des fins d'indexation nécessite une étape préalable de structuration de ce flux, basée sur une segmentation en composantes élémentaires puis en sujets spécifiques au sein des composantes. Typiquement, analyser un flux sonore diffusé à la radio ou à la télévision requiert la détection des différentes émissions puis, lorsque cela a un sens, une segmentation de chaque émission selon les sujets abordés. Cette dernière étape est particulièrement importante pour les informations (journaux télévisés ou flash d'informations) qui abordent différents sujets liés à l'actualité du moment (Allan *et al.*, 1998) : leur indexation ou leur structuration nécessite leur découpage en titres, reportages et transitions.

Pour mener à bien cette segmentation en sujets, on utilise classiquement des méthodes développées en traitement automatique des langues pour la segmentation thématique de corpus écrits, que l'on applique sur la transcription automatique de la parole des émissions, considérée comme un texte à part entière. La majorité de ces méthodes se base sur la notion de cohésion lexicale qui mesure à quel point l'usage fait du vocabulaire change entre deux thèmes. L'application directe de ces techniques dans le cadre d'une segmentation en sujets des informations pose toutefois plusieurs problèmes. Tout d'abord, la transcription automatique présente des caractéristiques différentes des textes écrits. En plus des erreurs de transcription — entre 10 % et 20 % des mots des émissions d'information sont généralement mal reconnus avec les systèmes de transcription actuels —, le texte produit automatiquement ne possède pas de marques de structure syntaxique fortes : notions de phrase et paragraphe non définies, pas de ponctuation... En revanche, on peut associer à la transcription des informations d'ordre acoustique comme la durée des pauses ou les changements de locuteurs. Par ailleurs, la nature même des documents implique certaines difficultés. Le temps alloué à chaque sujet varie fortement et peut être extrêmement court, ce qui rend les mesures de cohésion lexicale peu efficaces. Si les reportages y ont en général une durée conséquente, des nouvelles brèves émaillent aussi leur contenu. Les titres du journal constituent un cas typique où un sujet est abordé en quelques mots.

Ce dernier exemple soulève d'ailleurs deux questions fondamentales, fortement liées, à savoir d'une part la relation entre thème et sujet d'actualité et, d'autre part, la granularité visée de la segmentation. L'utilisation de techniques de segmentation thématique pour détecter les changements de sujets se justifie par le fait que, dans la plupart des cas, deux sujets consécutifs ne traitent pas du même thème. Cependant, les notions de sujet et de thème, bien que proches, ne sont pas strictement équivalentes. Le cas des titres est emblématique : faut-il considérer les titres dans leur ensemble ou bien chaque titre comme un sujet ? Du point de vue de la structure de l'émission, le premier choix peut paraître plus pertinent, alors que c'est l'inverse si l'on considère la question d'un point de vue du thème. Par ailleurs, la notion de thème n'est elle-même pas clairement définie et sujette à interprétation quant à la granularité envisagée. Un reportage traitant de deux aspects proches, par exemple le clan Saddam Hussein et le rôle de l'ONU en Irak, doit-il être considéré comme une unique entité sur la guerre en Irak ou comme deux entités ? À défaut d'avoir une solution tranchée pour ces deux questions, nous prenons le parti ici de nous baser sur une méthode de type segmentation thématique pour notre problème de découpage en sujets journalistiques et nous laissons, à l'instar de (Rossignol & Sébillot, 2005), la granularité des thèmes émerger des données.

Pour répondre à notre tâche de segmentation de transcriptions automatiques d'émissions d'actualités, nous décrivons, dans cet article, une extension du modèle de cohésion lexicale de Utiyama et Isahara (Utiyama & Isahara, 2001), modèle qui a fait ses preuves pour la segmenta-

tion thématique de l'écrit. Nous tirons profit du cadre probabiliste de ce modèle pour proposer une technique permettant d'y combiner différents types d'information, à savoir lexicale, syntaxique et acoustique. Cette combinaison inédite d'indices nous permet d'obtenir un découpage en sujets performant, sans faire d'hypothèses *a priori* sur les sujets présents, tout en étant robuste aux erreurs de transcription et à la durée potentiellement courte des segments.

Dans la suite, nous passons tout d'abord en revue les techniques classiquement utilisées pour détecter des changements de thèmes, avant de présenter en détail le modèle de cohésion lexicale utilisé et l'extension proposée aux autres sources d'information. Nous décrivons ensuite notre cadre expérimental et notre méthodologie de mesure des performances avant de présenter des résultats expérimentaux sur le corpus ESTER (Galliano *et al.*, 2006).

2 Techniques de segmentation thématique

La segmentation thématique consiste à localiser dans un document les points de changement de thèmes, de manière à obtenir des entités homogènes du point de vue du thème. Cette problématique a principalement été étudiée sur des données textuelles dans différents contextes applicatifs comme la recherche d'information ou le résumé automatique. Dans le cadre des documents écrits, les indices permettant de détecter ces points de changement thématique sont de plusieurs natures. Tout d'abord, au niveau lexical, un changement de thème s'accompagne en général d'une modification du vocabulaire utilisé. Dès lors, la notion de cohésion lexicale d'un segment, qui reflète l'homogénéité du vocabulaire — homogénéité « graphique » (répétitions) mais également sémantique, les mêmes lexies ou des lexies d'un même paradigme étant employées conjointement et différemment de leur distribution dans les segments adjacents —, est à la base de nombreuses méthodes que nous discutons par la suite. Au-delà de la mise au jour de ruptures de similarité lexicale entre portions consécutives, la segmentation thématique peut également s'appuyer sur le repérage de marqueurs discursifs, tant de continuité (« *de plus* ») que de changement (« *et maintenant* »), mais également sur des informations de type syntaxique telles la présence de pronoms référant des éléments antérieurs et suggérant la continuité d'un thème.

Dans le cas de documents sonores ou multimédias contenant de la parole, d'autres indices que ceux liés au langage naturel peuvent apporter une information sur les frontières de segments. Par exemple, lors des informations, les changements de locuteurs, la présence de musique et, dans le cas de la télévision, les changements de plans ou l'apparition du plateau sont des indices forts marquant un changement de sujet (Tür *et al.*, 2001; Galley *et al.*, 2003). De manière moins spécifique au type de documents étudiés, la prosodie fournit de précieux indices qui, en dehors de la durée des pauses silencieuses, restent peu utilisés du fait de la difficulté à les mesurer de façon fiable (Passonneau & Litman, 1997; Tür *et al.*, 2001).

Concrètement, un ou plusieurs des indices que nous venons de mentionner sont exploités par les diverses méthodes de segmentation thématique de la littérature. Lorsque les thèmes susceptibles d'être abordés sont connus à l'avance, il est possible de s'appuyer sur des modèles de langue thématiques pour la segmentation et la caractérisation des thèmes. Cette nécessité de définir *a priori* les thèmes limite toutefois les applications possibles de ces approches et les rend difficilement exploitables dans le cadre de la segmentation en sujets des informations. En l'absence de ces connaissances *a priori*, de nombreux travaux fondent leur repérage de ruptures thématiques sur la cohésion lexicale, et plus particulièrement sur la détection de minima locaux de similarité lexicale entre parties de textes consécutives. Outre le choix de la représentation des mots

par des lexies, des lemmes ou des racines, un aspect délicat de ces méthodes reste la définition des unités élémentaires pour la comparaison de la distribution lexicale. Dans le cas de textes structurés, le paragraphe peut être utilisé comme unité élémentaire, une mesure de rupture thématique étant effectué entre paragraphes successifs. À l'opposé, la technique peut se baser sur des fenêtres d'analyse contenant un nombre fixe d'unités (Hearst, 1997), ces dernières pouvant être des mots, des phrases ou encore des segments découpés lors de l'étape de la transcription automatique dans le cas de documents oraux. Ces méthodes nécessitent des fenêtres d'analyse suffisamment étendues ou l'apport de connaissances extérieures pour évaluer correctement les zones de ruptures de cohésion. Utiyama et Isahara proposent une approche originale dans laquelle l'ensemble des segmentations possibles est considéré, le problème de segmentation thématique consistant à trouver celle conduisant aux segments les plus homogènes (Utiyama & Isahara, 2001). À l'inverse des techniques exploitant une mesure de cohésion lexicale entre segments successifs, cette approche ne se base que sur la mesure de la cohésion au sein d'un segment. Lorsque des indices autres que lexicaux sont disponibles, ces différentes méthodes à base de cohésion lexicale peuvent être étendues pour les prendre en compte. Diverses techniques telles les arbres de décision (Tür *et al.*, 2001) ou les modèles de maximum d'entropie (Beeferman *et al.*, 1999), ont été ainsi étudiées pour mélanger des informations lexicales et acoustiques.

Notre cadre applicatif d'émissions d'actualités implique le repérage de segments de longueurs très variables, possiblement très courts, potentiellement problématiques pour la plupart des techniques « standards » de cohésion lexicale s'appuyant sur des fenêtres de taille fixe. Nous avons choisi d'exploiter pour l'oral la méthode de cohésion lexicale de Utiyama et Isahara, et de l'étendre en tirant profit de sa flexibilité, de ses bonnes performances et de son potentiel — non encore exploré — pour l'intégration d'informations syntaxiques et acoustiques.

3 Un modèle statistique multi-sources

Nous rappelons tout d'abord le principe proposé par Utiyama et Isahara avant de décrire une extension de ce formalisme.

3.1 Modèle statistique basé sur la cohésion lexicale

L'idée de la méthode de Utiyama et Isahara est de rechercher la segmentation qui conduit aux segments les plus homogènes sur le plan lexical tout en respectant une distribution *a priori* de la longueur des segments. La méthode se place dans un cadre probabiliste et consiste à trouver la meilleure segmentation $\widehat{S}$ d'une séquence de l unités élémentaires (mots, phrases...) $W = W_1^l$, parmi toutes les segmentations possibles. La loi de Laplace est utilisée pour modéliser la cohésion lexicale des segments, ces derniers étant vus comme des « sacs de mots ». De manière formelle, la meilleure segmentation est donnée par le critère du maximum *a posteriori*, soit, par application de la règle de Bayes,

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

En supposant d'une part que chaque segment forme une unité indépendante du reste du texte et, d'autre part, que les mots au sein d'un segment sont indépendants — ce qui revient à représenter un segment comme un « sac de mots » —, la probabilité d'un texte W pour une segmentation

$S = S_1^m$ est alors 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 dans le segment S_i et $w_j^{(i)}$ le j-ème mot de S_i. La probabilité $P[w_j^{(i)}|S_i]$ est fournie par une loi de Laplace dont les paramètres sont estimés sur S_i, soit

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

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 W. En d'autres termes, on mesure à quel point un modèle de la distribution des mots appris sur un segment S_i permet de prédire les mots de ce dernier. Intuitivement, cette probabilité favorise les segments homogènes du point de vue lexical puisqu'elle augmente lorsque les mots apparaissent plusieurs fois et qu'elle diminue si beaucoup de mots sont différents.

La mesure de cohésion lexicale précédente est complétée par une distribution *a priori* de la durée des segments, donnée en l'absence de connaissances explicites sur les documents à traiter, par $P[S_1^m] = n^{-m}$ où n est le nombre total de mots. Sa valeur est élevée lorsque le nombre de segments est petit, ce qui compense le fait que les valeurs de $P[W|S]$ sont fortes lorsque le nombre de segments est grand.

Sur le plan de la mise en œuvre, cette méthode de segmentation peut être vue comme la recherche du meilleur chemin dans un graphe valué représentant l'ensemble des segmentations possibles. Chaque nœud du graphe modélise une frontière de segment entre deux unités élémentaires, un arc entre deux nœuds i et j représentant alors un segment regroupant les unités élémentaires W_{i+1} à W_j. La valeur associée à un arc entre deux nœuds i et j est donnée par

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

en notant $S_{i\to j}$ le segment correspondant à l'arc. Le facteur α est ajouté en pratique pour pondérer la probabilité *a priori* et permettre un contrôle de la taille moyenne des segments retournés.

3.2 Introduction d'informations syntaxiques et acoustiques

Nous étendons le modèle précédent de manière à prendre en compte de nouvelles sources d'information. Nous considérons ici deux nouvelles sources, à savoir la syntaxe et l'acoustique, sans toutefois limiter la généralisation du formalisme proposé. En notant A l'ensemble des informations acoustiques disponibles et M les étiquettes morpho-syntaxiques associées aux mots de W, le critère à optimiser pour la segmentation s'écrit sous la forme

$$\widehat{S} = P[W, A, M|S]P[S] = P[W|S]P[A|S]P[M|S]P[S] \ , \tag{5}$$

en supposant les sources d'information indépendantes.

Le calcul des probabilités $P[A|S]$ et $P[M|S]$ suit le même principe que nous illustrons ici pour A. Dans la mesure où $P[A|S]$ est proportionnel à $P[S|A]/P[S]$, déterminer $P[A|S]$ se résume au calcul de $P[S|A]$. Nous proposons d'utiliser les informations acoustiques ou syntaxiques

pour prédire la probabilité qu'une frontière de segment se trouve entre chacune des unités élémentaires. En notant B_i la variable aléatoire binaire telle que $B_i = 1$ s'il existe une frontière entre W_i et W_{i+1} et en supposant les B_i indépendants, nous obtenons

$$P[S|A] = \prod_{i=1}^{l} P[B_i|A] \ . \tag{6}$$

Dans les expériences présentées par la suite, nous utilisons un arbre de décision pour le calcul des probabilités $P[B_i|A]$ à partir des caractéristiques acoustiques au voisinage de la frontière considérée. Pour les informations syntaxiques, un modèle N-gramme caché (Stolcke *et al.*, 1998) est utilisé pour calculer la probabilité d'une frontière $P[B_i|M]$ entre chaque unité élémentaire.

Si l'hypothèse d'indépendance entre les sources d'information permet d'employer indépendamment un modèle pour chacune d'elles dans l'équation (5), elle est quelque peu réductrice. En particulier, les informations morpho-syntaxiques et les mots sont fortement liés. Cependant, la représentation de W sous forme de sac de mots est relativement distincte de la notion de séquence d'étiquettes morpho-syntaxiques considérée pour M, justifiant ainsi l'hypothèse faite. Dans le cas de deux sources d'information fortement corrélées, par exemple X et Y, il est toujours possible de les intégrer conjointement en calculant par une méthode appropriée les probabilités $P[B_i|X, Y]$.

En pratique, les informations acoustiques et syntaxiques sont utilisées pour modifier les valeurs associées aux arcs dans le graphe des segmentations possibles. La valeur de l'arc $i \rightarrow j$ est alors donnée par

$$\begin{aligned} v(i,j) &= \sum_{k=i+1}^{j} \ln(P[W_k|S_{ij}]) + \beta_{\mathrm{A}} \left(\ln(P[B_j = 1|A]) + \sum_{k=i+1}^{j-1} \ln(P[B_k = 0|A]) \right) \\ &+ \beta_{\mathrm{M}} \left(\ln(P[B_j = 1|M]) + \sum_{k=i+1}^{j-1} \ln(P[B_k = 0|M]) \right) - \alpha \ln(n) \ , \end{aligned} \tag{7}$$

les paramètres β_{A} et β_{M} ayant pour but de pondérer l'apport de chacune des sources d'information. Comme précédemment, le paramètre α permet de contrôler la taille moyenne des segments retournés en donnant plus ou moins d'importance à la distribution *a priori* de la segmentation.

4 Conditions expérimentales

Les expériences présentées dans cet article ont été menées sur les journaux de France Inter et de France Info du corpus de développement ESTER (Galliano *et al.*, 2006), soit quatre journaux d'une heure chacun enregistrés le même jour. Deux des journaux, un par radio, sont utilisés comme données de développement pour régler les paramètres du modèle tandis que les deux autres sont utilisés comme données de test. Nous avons établi une segmentation de référence des quatre émissions considérées en distinguant cinq types de segments, correspondant à la structure de montage des émissions : les titres, les reportages, les brèves, les segments de remplissage et les publicités. Si les titres du journal forment un ensemble sur le plan de la structure, chaque titre aborde un sujet différent. Nous avons choisi d'étiqueter chacun comme un segment, suivant ainsi un critère thématique, même si ces titres sont considérés différemment des autres sections lors de l'évaluation des performances comme discuté ultérieurement. Les reportages correspondent au développement des sujets d'actualités. Ils sont souvent longs, parfois entrecoupés

d'*interviews* ou encore d'interventions de spécialistes. Cependant, plusieurs sujets peuvent être abordés dans un reportage. Par exemple, un reportage sur la guerre en Irak aborde la traque des proches de Saddam Hussein, le rôle de l'ONU en Irak et enfin les conditions sanitaires à l'issue des combats. Comme pour les titres, nous avons choisi d'avoir un segment pour chacun des sujets abordés dans un reportage. Les brèves correspondent à des nouvelles courtes, généralement diffusées entre deux reportages. La difficulté de ce type de segment tient d'une part à leur durée extrêmement courte, de l'ordre d'une phrase ou deux, et, d'autre part, à l'agrégation possible de plusieurs nouvelles brèves dans une même phrase. Les segments de remplissage correspondent à des annonces de la radio pour donner l'heure, rappeler le nom des journalistes ou annoncer une rubrique (la météo, la bourse...). De par leur nature, ces segments ne sont associés à aucun sujet. En revanche, ils marquent la plupart du temps un changement de segment. Finalement, chaque publicité est annotée comme un segment à part entière.

La transcription automatique requiert une étape préalable de segmentation du flux sonore en pseudo-phrases, appelés groupes de souffle, chaque groupe de souffle correspondant à un énoncé entre deux respirations. Nous considérons dans cet article une segmentation manuelle en groupes de souffle, telle qu'établie selon les conventions de transcription du corpus ESTER, afin de pouvoir mesurer la dégradation de performance due à la transcription automatique. Les différentes étapes du processus de transcription automatique des journaux radiophoniques sont décrites en détail dans (Huet *et al.*, 2007). Pour les quatre journaux considérés, le taux d'erreur sur les mots est d'environ 20 %. Pour le modèle de cohésion lexicale employé dans la segmentation en sujets, nous ne conservons que les noms communs, les noms propres et les adjectifs après lemmatisation. Le taux d'erreur sur les lemmes conservés est d'environ 17 % sur les quatre heures d'émissions.

L'évaluation des performances se fait sur la base du nombre de frontières entre groupes de souffle détectées correctement, les résultats étant donnés en termes de rappel et précision ainsi que selon la métrique *WindowDiff* (Pevzner & Hearst, 2002). En raison des difficultés évoquées précédemment concernant les titres et la granularité considérée pour la segmentation des reportages, la notion de frontières correctes n'est pas immédiate. Nous distinguons deux types de frontières selon qu'elles sont considérées comme facultatives ou obligatoires. Les frontières entre titres successifs ainsi qu'entre sujets proches au sein d'un même reportage sont considérées comme facultatives. Si de telles frontières n'entrent pas dans la comptabilisation des erreurs, les détecter ne constitue pas pour autant une erreur d'insertion. Enfin, les segment de remplissage posent un problème particulier car ils ne sont pas liés à un sujet particulier. La détection de tels segments nécessite des méthodes *ad hoc* et peu portables, par exemple basées sur une liste de marqueurs pré-établis, que nous ne souhaitons pas considérer dans ce travail par souci de généricité. Nous adoptons plutôt la notion de frontière floue en acceptant comme correcte non seulement la détection des deux frontières du segment de remplissage mais aussi l'agrégation de ce dernier au segment précédent ou suivant. La métrique *WindowDiff* a été adaptée pour prendre en compte les considérations que nous venons d'énumérer.

5 Résultats

Nous présentons tout d'abord quelques résultats basés sur le seul modèle de cohésion lexicale. Le poids α de la distribution *a priori* dans l'équation (7) permet de faire varier la durée moyenne des segmentations et donc d'atteindre différents compromis entre rappel et précision. Nous avons réglé ce poids de manière à minimiser le critère *WindowDiff* sur le corpus de déve-

loppement. Sur ce dernier, les valeurs de rappel et de précision sur les frontières de segments, après optimisation de α, sont respectivement de 37,5 % et 54,1 % pour une durée moyenne des segments de 26,5 groupes de souffle (*WindowDiff*=0,110). Ces résultats mettent en évidence d'une part les performances limitées de la seule cohésion lexicale pour segmenter des journaux radiophoniques et, d'autre part, la dégradation de performances liée aux erreurs de transcription. En effet, en remplacant la transcription automatique par la référence, nous obtenons un rappel de 43,8 % pour une précision de 54,5 %, correspondant à un critère *WindowDiff* de 0,105. Les faibles valeurs de rappel et de précision, par opposition aux relativement bonnes performances obtenues selon le critère *WindowDiff*, montrent que nombre d'erreurs sont dues à une légère imprécision sur la position des frontières. En effet, en acceptant une tolérance d'un groupe de souffle sur la position des frontières, on obtient alors un rappel de 58,8 % et une précision de 79,7% pour la transcription automatique, ce qui illustre bien le problème de précision sur la position des frontières de segments.

Pour améliorer les premiers résultats basés sur la seule cohésion lexicale, nous étudions l'apport des informations syntaxiques et acoustiques. Les modèles utilisés pour prédire la probabilité d'une frontière à partir de l'une ou l'autre des sources d'information ont été appris sur les annotations de référence des 80 heures du corpus d'apprentissage ESTER en se basant sur la segmentation en sujets présente dans le corpus. Bien que sujette à caution sur certains points[1], cette dernière indique clairement les frontières de reportages et les alternances reportages, brèves et segments de remplissage. Nous utilisons un arbre de décision pour calculer la probabilité d'une frontière de segment entre groupes de souffle sur la base des informations acoustiques. Les indices acoustiques considérés sont la présence d'un *jingle* à la frontière, les durées du *jingle* en cours respectivement avant et après la fin du groupe de souffle, l'alternance ou pas de locuteurs homme/femme[2] et la durée des pauses en ne considérant que les pauses supérieures à 1,5 secondes. De manière surprenante, la durée des pauses silencieuses n'a pas été retenu comme critère dans l'arbre de décision pour prédire la présence d'une frontière de segment. En effet, nous avons constaté que de longues plages de silence intervenaient lors des interviews, expliquant ainsi la faible corrélation entre durée des pauses et frontière thématique. L'alternance de locuteurs homme/femme a été jugée comme le critère le plus pertinent, suivie par la présence de musique. Les informations syntaxiques sont quant à elles prises en compte par l'intermédiaire d'un modèle N-gramme caché, avec N=6, sur les étiquettes morpho-syntaxiques.

Les résultats obtenus sur le corpus de développement, en tenant compte des informations lexicales, syntaxiques et acoustiques, sont donnés dans la figure 1. Pour l'ensemble des courbes présentées, les poids β_{A} et β_{M} ont été déterminés de manière à minimiser le critère *WindowDiff* sur les données de développement. Ces résultats montrent clairement l'apport des informations syntaxiques par rapport au modèle initial de cohésion lexicale. Bien que moindre, l'apport des informations acoustiques est également mis en évidence. Ce dernier résultat s'explique en partie par le fait que les indices acoustiques permettent surtout d'éviter les frontières considérées comme optionnelles, notamment entre titres. Cette amélioration n'a qu'un faible impact sur la mesure du rappel et de la précision en raison de la métrique utilisée pour laquelle la détection des frontières dites optionnelles n'est pas considérée comme une erreur. Enfin, la combinaison de l'ensemble des sources d'information permet d'obtenir les meilleurs résultats, ce qui est

[1] À l'inverse des choix que nous avons effectués, les titres sont considérés comme un unique segment. Il en est de même pour les reportages abordant plusieurs sujets.

[2] L'alternance de locuteurs homme/femme a été préférée aux changements de locuteurs pour deux raisons : d'une part, la détection automatique des changements de sexe du locuteur est déjà un indice intéressant pour le passage d'un sujet à un autre. D'autre part, cette détection est plus fiable que celle des locuteurs.

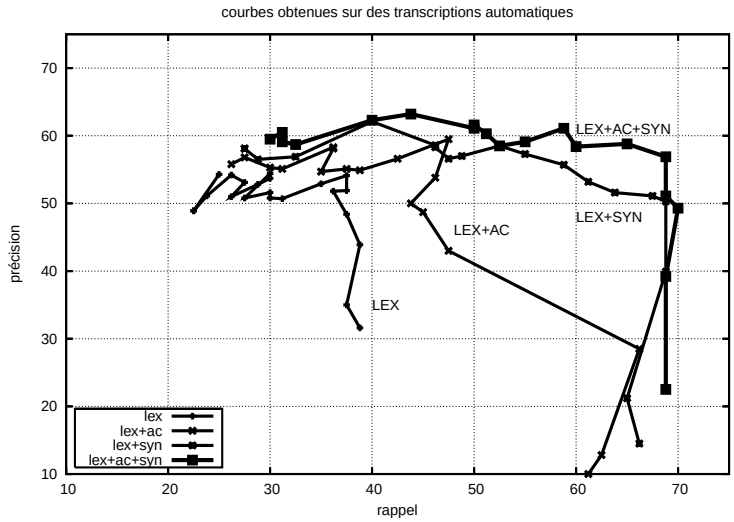

FIG. 1 – Rappel et précision pour la cohésion lexicale seule (LEX), avec l'acoustique (LEX+AC), la syntaxe (LEX+SYN) ou encore les deux (LEX+AC+SYN). Les courbes sont obtenues en faisant varier les valeurs des pondérations sur les deux émissions du corpus de développement.

	longueur moyenne	pureté segment	rappel	précision	F	wDiff
lex	23,8	0,76	45,2	51,5	48,1	0,098
lex+ac	38,5	0,69	43,5	70,0	53,7	0,121
lex+syn	29,9	0,73	45,2	61,5	52,1	0,097
lex+ac+syn	20,5	0,81	56,5	62,3	59,3	0,090

TAB. 1 – Résultats sur les deux émissions de test. Le nombre moyen de groupes de souffle par segment selon la segmentation de référence est sur ce corpus de 24,7.

confirmé sur les données de test (*cf.* tableau 1). L'utilisation d'indices acoustiques permet d'obtenir des segments plus longs, augmentant ainsi grandement la précision pour une faible baisse du rappel ; ceci se fait au détriment de la pureté qui mesure la part des segments qui ne traitent que d'un seul sujet. Les informations syntaxiques permettent également une augmentation de la précision sans toutefois modifier le rappel ni outre mesure la longueur moyenne des segments. L'utilisation conjointe des indices acoustiques et syntaxiques permet une amélioration à la fois de la précision et du rappel, ce qui se traduit par une amélioration sensible du critère *WindowDiff* par rapport à la seule cohésion lexicale. La pureté des segments obtenus est également grandement améliorée.

6 Conclusion

Nous avons proposé dans cet article un modèle statistique permettant la prise en compte simultanée d'indices lexicaux, syntaxiques et acoustiques pour la segmentation en sujets de journaux radiophoniques. Les résultats expérimentaux ont permis de mettre en évidence que les limites du modèle de cohésion lexicale pour le type de documents utilisés ainsi que sa sensibilité aux erreurs de transcription sont compensées par les informations syntaxiques et acoustiques.

Bien qu'appliqué à deux sources de connaissances, notre modèle offre un cadre générique permettant d'accueillir d'autres indices. Il convient de noter toutefois que subsiste un aspect supervisé qui pourrait limiter la robustesse de la méthode par rapport à la diversité des types de documents : l'apprentissage des modèles de prédiction des frontières effectué sur un corpus segmenté manuellement. Si les indices considérés ici sont suffisamment généraux pour donner à penser que l'extension à d'autres documents est possible, une vérification est toutefois nécessaire. C'est d'ailleurs l'une des premières perspectives du travail présenté. Une autre piste consiste à étudier l'apport de connaissances sémantiques, syntagmatiques et paradigmatiques, à la segmentation, connaissances que notre modèle est à même de représenter. Ces informations permettront entre autres de compenser la faible taille de certains segments. Enfin, il s'avère important d'atténuer l'impact des erreurs de transcription, par exemple en ne se limitant pas à la seule transcription mais en exploitant d'autres représentations plus riches fournies par le système de reconnaissance automatique (graphes de mots, indices de confiance...).

Références

ALLAN J., CARBONELL J., DODDINGTON G., YAMRON J. & YANG Y. (1998). Topic detection and tracking pilot study final report. In *Proc. DARPA Broadcast News Transcription and Understanding Workshop*.

BEEFERMAN D., BERGER A. & LAFFERTY J. (1999). Statistical models for text segmentation. *Machine Learning*, **34**(1-3), 177–210.

GALLEY M., MCKEOWN K., FOSLER-LUSSIER E. & JING H. (2003). Discourse segmentation of multi-party conversation. In *Proc. Association for Computational Linguistics*.

GALLIANO S., GEOFFROIS E., BONASTRE J.-F., GRAVIER G., MOSTEFA D. & CHOUKRI K. (2006). Corpus description of the ESTER evaluation campaign for the rich transcription of French broadcast news. In *Proc. Language Resources and Evaluation Conference*.

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

HUET S., GRAVIER G. & SÉBILLOT P. (2007). Morphosyntactic processing of N-best lists for improved recognition and confidence measure computation. In *Proc. European Conf. on Speech Communication and Technology*.

PASSONNEAU R. J. & LITMAN D. J. (1997). Discourse segmentation by human and automated means. *Computational Linguistics*, **23**(1), 103–139.

PEVZNER L. & HEARST M. A. (2002). A critique and improvement of an evaluation metric for text segmentation. *Computational Linguistics*, **28**(1), 19–36.

ROSSIGNOL M. & SÉBILLOT P. (2005). Combining statistical data analysis techniques to extract topical keyword classes from corpora. *Intelligent Data Analysis*, **9**(1), 105–127.

STOLCKE A., SHRIBERG E., BATES R., OSTENDORF M., HAKKANI D., PLAUCHE M., TÜR G. & LU Y. (1998). Automatic detection of sentence boundaries and disfluencies based on recognized words. In *Proc. Intl. Conf. on Spoken Language Processing*.

TÜR G., HAKKANI-TÜR D., STOLCKE A. & SHRIBERG E. (2001). Integrating prosodic and lexical cues for automatic topic segmentation. *Computational Linguistics*, **21**(1), 31–57.

UTIYAMA M. & ISAHARA H. (2001). A statistical model for domain-independent text segmentation. In *Proc. Association for Computational Linguistics*.