Reconnaissance thématique à partir de textes dictés et Adaptation dynamique de modèles de langages thématiques

Brigitte Bigi, Renato de Mori, Spriet Thierry

Reconnaissance thématique à partir de textes dictés et Adaptation dynamique de modèles de langage thématiques

Brigitte Bigi, Renato De Mori, Thierry Spriet

Laboratoire d'Informatique d'Avignon
L. I. A. - CERI BP 1228 - 84911 Avignon Cedex 9 - FRANCE
Tél. : ++33 (0)4 90 84 35 36 - Fax : ++33 (0)4 90 84 35 01
Mél : {brigitte.bigi,renato.demori,thierry.spriet}@lia.univ-avignon.fr

Résumé

A robust strategy for dynamic language model selection, based on topic recognition and switching between topic models, is proposed. It is effective because it relies on a small set of well trained topic-dependent language models and on reliable topic recognition. By using perplexity as a performance measure of the LM switching model, a tangible reduction is observed with respect to the use of a single, general, static LM. Different methods are proposed for topic shift detection. Experimental results show that different strategies for topic shift detection have to be used depending on whether high recall or high precision are sought.

Présentation

Dans le cadre d'une amélioration des performances des systèmes de Reconnaissance Automatique de la Parole (RAP), nos travaux visent l'adaptation dynamique de leur composante linguistique. Cette adaptation est réalisée en fonction des thèmes identifiés dynamiquement lors de la dictée. Le principe général de cette approche est illustré par la figure 1. Il consiste à utiliser un modèle généraliste au début de la reconnaissance afin d'initialiser les processus de classification thématique. Par la suite, on adapte le modèle de langage (ML) en fonction du résultat de la classification. Le problème de ce type de modèles est qu'ils nécessitent une grande quantité de *corpus segmenté en thèmes*. Cet article présente un ensemble de solutions possibles qui consistent à segmenter automatique des données qui pourront, par la suite, être utilisées pour l'apprentissage.

En section 1, nous montrons brièvement les méthodes développées pour obtenir une classification thématique sur des textes écrits, puis sur des textes dictés à un système de reconnaissance de la parole. De plus, nous montrerons le potentiel de reconnaissance des modèles de langages thématiques en évaluant le gain de perplexité qu'ils peuvent engendrer. En section 2, nous développons différentes méthodes de segmentation qui génèrent les ruptures thématiques.

1. Classification thématique

La classification thématique est un processus appliqué à un texte et dont le résultat est l'assignation d'un label thématique parmi une liste prédéfinie de labels possibles. Ce travail, déjà présenté en [1], propose l'utilisation de deux ML : un ensemble d'unigrammes thématiques, et un modèle basé sur une mémoire cache [3]. Ce dernier calcule, pour chaque thème, la distance entre une distribution de mots-clés thématiques et le contenu de la mémoire cache. A partir de ces distances, nous obtenons des probabilités associées à chaque thème pour le texte.

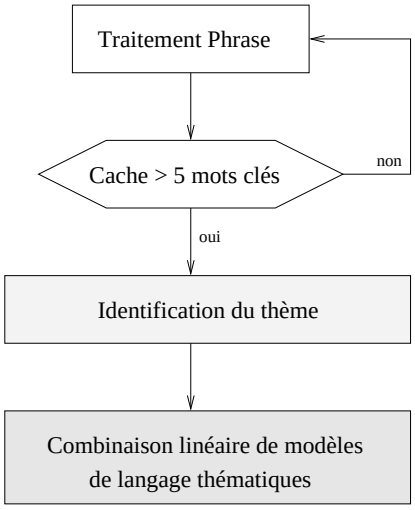


FIG. 1 – Processus d'Adaptation dynamique de modèles de langage thématiques

Les expériences ont été réalisées sur des articles du journal "Le Monde" de 1987 à 1991, avec un vocabulaire de plus de 500 000 entités lexicales. Les sept thèmes retenus sont issus des secteurs rédactionels du journal : *Etranger, Histoire, Sciences, Sports, Economie, Culture, Politique*. Les résultats donnent un taux de classification thématique de plus de 80 % sur un corpus de test composé de 1021 paragraphes.

1.1. Reconnaissance thématique sur corpus dicté

Pour évaluer le taux de classification thématique de textes dictés, nous avons comparé le label assigné par notre système au texte original à celui assigné au texte dicté (sortie du système de reconnaissance). Le corpus de test composé de 97 paragraphes (18000 mots) a été dicté au système *ViaVoice 98* d'IBM, un système dépendant du locuteur et à grand vocabulaire. Quatre locuteurs (1 femme, 3 hommes) ont participé à sa constitution. Nous avons obtenus un taux d'erreurs d'environ 35 % sur ce corpus.

Nous obtenons 73 paragraphes correctement étiquetés contre 77 sur les textes de référence correspondants, ce qui signifie que les mots clés sont apparemment bien reconnus. Ce résultat confirme que nos travaux de classification thématique sur l'écrit peuvent être réutilisés dans un système de dictée.

1.2. Bigramme thématique

L'expérience consiste à utiliser un modèle de langage thématique sur un corpus de ce thème (ML_t) et d'en comparer la perplexité avec celle d'un modèle de langage général (ML_g). Pour ce faire, on crée un ensemble de bigrammes thématiques et un bigramme général, tous de même vocabulaire (V=10000).

Afin que les modèles thématiques puissent être estimés avec peu de corpus d'apprentissage spécifique au domaine, nous utilisons la combinaison linéaire de chaque bigramme thématique avec le bigramme général :

$$P(w_i|w_j) = \lambda P_g(w_i|w_j) + (1-\lambda)P_t(w_i|w_j)$$

Le coefficient λ, appliqué à ML_g et estimé empiriquement, reflète la qualité du bigramme thématique par rapport à ML_g. Le calcul de la perplexité de ML_g et celui de chaque modèle combiné est réalisé pour chaque thème. Ceux-ci ont été effectués avec le toolkit v2 du CMU ([6]). Les résultats sont reportés en table 1. La deuxième colonne indique la quantité de données utilisées lors de l'apprentissage des modèles thématiques ; le ML_g, quant à lui, a nécessité un autre corpus de 32,3 M de mots. Le thème *histoire* n'ayant pu produire un modèle thématique cohérent, l'estimation du λ a rejeté l'utilisation de ce thème. Un gain de l'ordre de 8,7 % est observé sur les 6 autres thèmes.

TAB. 1 – Comparaisons de perplexités des modèles bigrammes combinés, et du modèle général

Thème	Taille (mots)	λ	PP ML combiné	PP ML général	Gain
Etranger	15,1	0,3	135,9	149	8,8 %
Histoire	0,6	1	-	191,2	-
Science	2	0,6	156,3	174,2	10,3 %
Sport	0,2	0,7	161,5	179,5	10 %
Economie	10,4	0,3	133,9	144,9	7,6 %
Culture	15,8	0,3	160,6	177	9,3 %
Politique	10,1	0,4	138,6	148,4	6,6 %

2. Segmentation thématique

La segmentation thématique a pour but de déterminer automatiquement les frontières thématiques des segments qui composent un document. Les résultats s'expriment sous la forme de taux de rappel et précision, avec :

$$rappel = \frac{Nb\ de\ ruptures\ correctes\ trouvees}{Nb\ de\ ruptures\ a\ trouver}$$

$$precision = \frac{Nb\ de\ ruptures\ correctes\ trouvees}{Nb\ de\ ruptures\ totales\ trouvees}$$

Plusieurs stratégies pour repérer les ruptures de thèmes et pour les sélectionner sont possibles. Une partie de ces travaux est présentée dans [2].

Le corpus de test est composé de 1393 documents générés automatiquement de telle sorte que chacun d'entre-eux est constitué de 3 paragraphes tirés aléatoirement dans nos corpus thématiques. Dans la mesure où la taille des paragraphes qui constituent les documents peut avoir une conséquence importante sur la qualité des résultats, aucune contrainte n'a été appliquée sur celle-ci.

2.1. Première phase de détection des ruptures candidates

Deux approches ont été étudiées. La première reprend les travaux développés pour la classification thématique, et utilise un modèle basé sur une mémoire cache. La seconde approche, place les candidats à intervalles réguliers sans tenir compte de la nature du texte.

Le modèle à base de mémoire cache

Comme décrit dans la section précédente, ce modèle est un outil performant pour la classification thématique, c'est pourquoi nous avons voulu l'utiliser dans le cadre de la segmentation. Le contenu de la mémoire cache est une représentation de l'historique puisqu'elle en conserve uniquement les 100 derniers mot clés. La mémoire cache est réinitialisée à chaque nouveau document. Nous utilisons la distance $d_j^*(i)$, distance de Kullback-Liebler normalisée évaluée entre le contenu de la mémoire cache et l'histogramme des mots clés du thème T_j. Pour la classification thématique, le thème assigné au paragraphe est celui dont la distance est la plus petite.

La distance $d_j^*(i)$ est la distance entre la mémoire cache et l'histogramme du thème T_j à la fin de la i-ème phrase du document. Nous nous intéressons à l'évolution de cette distance pour le meilleur thème, au sens de la classification. Cette variation s'exprime par :

$$\delta(i) = d_j^*(i) - d_j^*(i-1)$$

où j est le meilleur thème à la fin de la $(i-1)$ème phrase.

Nous proposons une rupture candidate à chaque variation importante de la distance, c'est à dire quand $\delta(i) > \theta$, où θ est un seuil déterminé expérimentalement.

Les résultats que l'on obtient avec cette méthode sont exprimés dans la table 2. On observe une valeur élevée de rappel, ce qui signifie que peu de frontières thématiques n'ont pas été détectées. Par contre, la faible valeur de précision indique un nombre important de fausses alarmes. Les différentes valeurs de θ ne font que faire varier proportionnellement les valeurs de rappel et de précision.

TAB. 2 – Repérage de ruptures thématiques par le modèle à base de mémoire cache

θ	0,001	0,0014	0,0018	0,002	0,003
Rappel	0,9091	0,8664	0,8152	0,789	0,612
Précision	0,1261	0,1747	0,2309	0,2608	0,4047

Méthode systématique

A fin de comparaison, nous avons aussi utilisé une méthode systématique. Elle place arbitrairement un candidat toutes les N phrases. Cette méthode permet avec $N = 1$ de trouver toutes les ruptures (rappel=1), avec la précision minimale.

Les résultats sont donnés dans la table 3. Comme on pouvait s'y attendre, on constate qu'ils sont moins bons que ceux de la méthode précédente.

TAB. 3 – Repérage des ruptures par le placement d'un candidat toutes les N phrases

N	1	5	10
Rappel	1	0,8391	0,6816
Précision	0,0249	0,1084	0,1805

2.2. Sélectionner les candidats

On définit un segment comme étant la portion de texte comprise entre deux ruptures candidates. Dans la première étape, on a exposé des méthodes pour repérer un ensemble de ruptures candidates qui peuvent représenter la frontière thématique entre deux segments de texte. On observe donc de forts taux de rappel, ce qui satisfait notre objectif. Dans cette deuxième étape, on développe plusieurs méthodes qui devront sélectionner les candidats en minimisant les fausses alarmes, sans trop perdre de la valeur de rappel.

Utilisation de la distance du modèle cache

En utilisant le modèle à base de mémoire cache pour leur classification, il est possible que deux segments successifs soient étiquetés avec le même thème. Les ruptures de thèmes sont alors définies lorsque deux labels thématiques différents sont observés dans deux segments adjacents. On peut noter que, comme les ruptures candidates sont obtenues avec des règles locales appliquées au contenu de la mémoire cache, quand on génère des segments de textes qui ne contiennent pas un nombre suffisant de mots la rupture candidate est ignorée. Ce nombre a été fixé empiriquement en fonction de la méthode utilisée.

TAB. 4 – Méthode de repérage par le modèle cache et sélection par le modèle cache

θ	0,001	0,0014	0,0018	0,002	0,003
Rappel	0,3857	0,3797	0,3715	0,3618	0,2851
Précision	0,6492	0,6794	0,7394	0,7543	0,8141

TAB. 5 – Méthode de repérage systématique avec sélection par le modèle cache

N	1	5	10
Rappel	0,3958	0,4706	0,3726
Précision	0,4235	0,5197	0,4911

Les résultats sont dans les tables 4 et 5 en fonction de la méthode de détection des candidats qui a été utilisée. On remarque l'importance de la méthode de repérage lors de la phase de sélection, puisque les candidats choisis par la méthode à base de mémoire cache donnent de meilleurs résultats qu'avec la méthode systématique. On le voit notamment en comparant les cas où $\theta = 0,0018$ et $N = 10$. Pour la même valeur de rappel, la valeur de précision est nettement meilleure par le repérage à base de mémoire cache.

Distance entre le contexte gauche et le contexte droit

Dans cette méthode, nous recherchons une séquence optimale de candidats, selon un critère d'optimisation. L'avantage essentiel que propose cette méthode est que, contrairement à la précédente, elle ne nécessite pas de connaissances *a priori* des thèmes.

C'est une programmation dynamique dont l'automate est celui de la figure 2. On note r, le cas où la rupture candidate est effective, et c le cas où c'est une continuité.

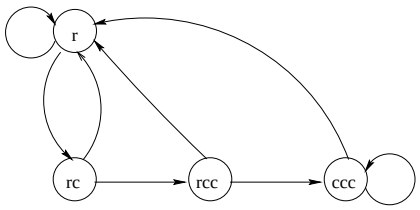


FIG. 2 – Automate de la méthode de sélection des candidats

Le treillis d'évaluation qui résulte de cet automate est présenté dans la figure 3. On note RP_i la rupture potentielle au i-ème segment. L'évaluation des

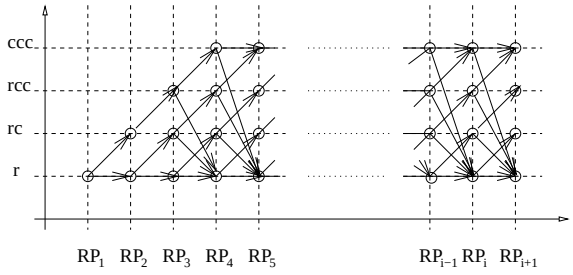


FIG. 3 – Treillis de la programmation dynamique selon l'automate de la figure 2

points de ce treillis s'effectue en 3 étapes. Dans un premier temps, on évalue la distance $d_i(G, D)$, distance de Kullback-Liebler entre le contexte gauche et le contexte droit, du i-ème candidat, où un nombre différent de segments peut être utilisé pour représenter le contexte gauche, en fonction de l'état pour lequel on calcule cette distance. Ceci est illustré par la figure 4. Ensuite, on analyse les variations de cette distance avec celles des états précédents potentiels :

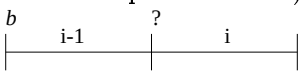
$$\Delta_i = d_i(G_i, D_i) - d_{i-1}(G_{i-1}, D_{i-1})$$

Enfin, on calcule les probabilités de continuité $P(c)$ et celles d'une rupture $P(r)$ telles que :

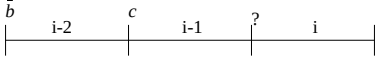
$$P(c) = \frac{\alpha}{1 + \exp^{-\Delta_i}}$$

$$P(r) = 1 - P(c)$$

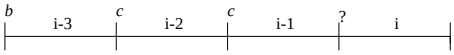
– état 1 : *r*. Le contexte gauche est représenté par un seul segment (i. e. la rupture candidate au $(i-1)$-ème segment est une rupture effective).



– état 2 : *rc*. Le contexte gauche contient les deux segments précédents.



– état 3 : *rcc*. Le contexte gauche contient les trois segments précédents.



– état 4 : *ccc*. Le contexte gauche contient les trois segments précédents.

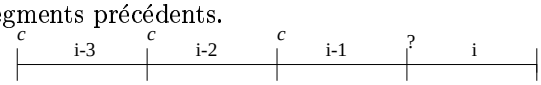


FIG. 4 – Différentes tailles de G, pour l'évaluation de la distance $d_i(G, D)$

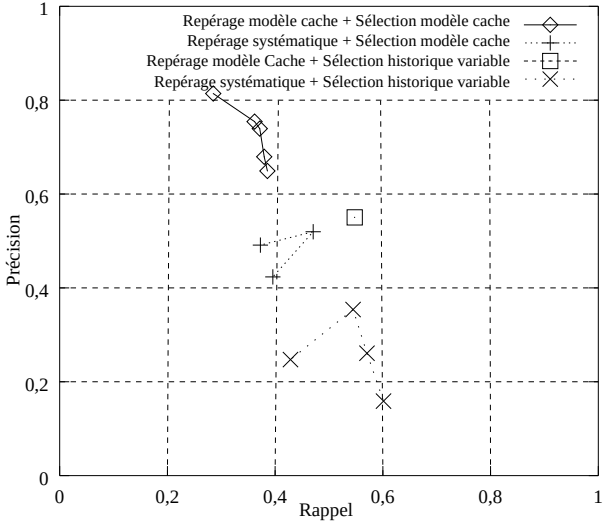


FIG. 5 – Résultats de la segmentation thématique (tables 4,5,6 et 7)

L'état précédent que l'on choisit est celui dont $P(c)$ est la plus grande. On remarque que dans le cas où $\Delta_i = 0$ et $\alpha = 1$, on aura $P(c) = P(r) = 0,5$. Les résultats de cette méthode de sélection sont en tables 6 et 7.

TAB. 6 – Méthode de repérage par le modèle cache ($\theta = 0,002$), et sélection par la méthode à historique variable

α	3
Rappel	0,5477
Précision	0,5502

TAB. 7 – Méthode de repérage systématique et sélection par la méthode à historique variable

N	5	5	5	10
α	1	2	3	1
Rappel	0,6012	0,5705	0,5447	0,4284
Précision	0,1587	0,2607	0,3539	0,2473

2.3. Synthèse des résultats

La figure 5 donne la courbe de rappel et précision que l'on obtient avec l'ensemble des différentes méthodes utilisées. Ces résultats montrent que différentes stratégies doivent être utilisées selon les valeurs de rappel ou de précision que l'on cherche à obtenir.

Perspectives

Dans cet article, on a montré que l'on peut déterminer rapidement le thème d'un paragraphe. On a vu aussi que les modèles de langage résultant de la combinaison linéaire de modèles thématiques et d'un modèle général peuvent apporter des gains substantiels de perplexité. Ce résultat pourra par la suite être validé en dictée réelle. On a ainsi tous les éléments pour réaliser une adaptation en fonction des thèmes identifiés dynamiquement lors de la dictée. Le problème à résoudre est de disposer d'un corpus suffisant pour apprendre les modèles de langage thématiques. C'est pourquoi, nous avons développé des méthodes de segmentation en thèmes. Les meilleures méthodes proposées utilisent le modèle cache qui nécessite des connaissances préalables sur les thèmes. Il serait préférable que la segmentation thématique puisse être obtenue sans nécessiter ces connaissances *a priori*. Une partie de la solution est proposée par la méthode de sélection à historique variable. La suite de ce travail est de trouver des méthodes efficaces pour repérer les candidats qui, elles-aussi, ne nécessiteront pas de connaissances préalables sur les thèmes.

Références

[1] B. Bigi, R. De Mori, M. El-Beze, T. Spriet A Fuzzy Decision Strategy for Topic Identification and Dynamic Selection of Language Models *Special Issue on Fuzzy Logic in Signal Processing*, Signal Processing Journal, volume 80, numero 6, 2000.

[2] B. Bigi, R. De Mori, M. El-Beze, T. Spriet Detecting topic shifts using a cache memory *5th International Conference on Spoken Language Processing*, ICSLP-98, Sydney, Australia

[3] R. Kuhn and R. De Mori. A cache-based natural language model for speech recognition. In *IEEE Trans. Pattern anal. Machine Intell*, PAMI-12(6), pp 570–582, 1990.

[4] H. Li and K. Yamamishi. Document classification using a finite mixture model. Proc. of the *Conference of the Association for Computational Linguistics*, pp 39–47, Madrid, Spain, 1997.

[5] Peskin, S. Conolly, L. Gillick, S. Lowe, D. McAllaster, V. van Mulbregt, and S. Wegmann. Improvements in switchboard recognition and topic identification. In *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 303–306, Atlanta GA, 1996.

[6] R. Rosenfeld. The CMU Statistical Language Modeling Toolkit, and its use in the 1994 ARPA CSR Evaluation" Proc. of the *ARPA Spoken Language Technology Workshop*, pp 47–50, Austin, Texas, 1995.