Combinaison de modèles de langage pour l'identification de thèmes

Brigitte Bigi, Renato De Mori, Marc El Bèze, Thierry Spriet

Combinaison de modèles de langages pour l'identification de thèmes

Brigitte Bigi, Renato De Mori, Marc El-Bèze, Thierry Spriet

L. I. A. , Université d'Avignon et des Pays de Vaucluse
CERI-IUP, BP 1228, 84911 Avignon Cedex 9 - France
Tél. : +33 09 90 84 35 00 - Fax : +33 09 90 84 35 01
{brigitte.bigi,renato.demori,marc.elbeze,thierry.spriet}@lia.univ-avignon.fr

Résumé

A new statistical method for Language Modeling and spoken document classification is proposed. It is based on a mixture of topic dependent probabilities. Each topic dependent probability is in turn a mixture of n-gram probabilities and the probability of Kullback-Lieber (KL) distances between key-word unigrams and distribution obtained from the content of a cache memory. Experimental result on topic classification using a corpus of 60 Mwords from the French newspaper *Le Monde* show the excellent performance of the cache memory and its complementary role in providing different statistics for the decision process.

Introduction

La classification en thèmes a suscité de nombreux travaux en recherche documentaire. Plusieurs méthodes ont été développées pour la classification de documents écrits. Récemment, des méthodes statistiques fondées sur les modèles unigrammes ont été proposées ([LIY97]).

L'utilisation de modèles de langage thématiques (ML) fait actuellement l'objet d'un intérêt croissant ([WRI96] ; [CAR96] ; [PES96]). Pour la transcription automatique de documents de parole ([WRI96] ; [CAR96] ; [PES96]), les mélanges de ce type de modèles sont souvent employées ([IMA97] ; [LIY97]).

En général, un mélange combine les différentes probabilités des ML, en calculant la probabilité d'un mot w_i sachant son histoire h_i de la façon suivante :

$$P(w_i|h_i) = \sum_{j=1}^{J} \lambda_j P_j(w_i|h_i) \qquad (1)$$

où $P_j(w_i|h_i)$ est la probabilité donnée par le j-ème ML et λ_j est un jeu de coefficients satisfaisant la contrainte :

$$\sum_{j=1}^{J} \lambda_j = 1$$

Il est alors possible de rendre λ_j dépendant du thème et égal aux scores qui sont utilisés pour l'identification de thèmes. En fait, $P(w_i|h_i)$ peut être exprimé comme suit :

$$\begin{aligned} P(w_i|h_i) &= \sum_{j=1}^{J} P(w_i, T_j|h_i) \\ &= \sum_{j=1}^{J} P(w_i|T_j, h_i)P(T_j|h_i) \end{aligned} \qquad (2)$$

où T_j est un thème. Le terme $P(w_i|T_j, h_i)$ est évalué avec un modèle n-gram qui dépend du thème. Ceci correspond à faire une approximation telle que h_i est représenté seulement par les $n-1$ mots qui précèdent w_i. Le second terme, $P(T_j|h_i)$, peut être calculé en utilisant une approximation différente pour h_i. Ce dernier peut, en effet, être utilisé pour moduler *dynamiquement* l'importance des éléments du mélange et pour l'identification des thèmes. Cette probabilité peut également être exprimée comme un mélange de probabilités obtenue avec différents modèles, et, plus important encore, elle peut varier dans le temps, au fur et à mesure de l'entrée de nouveaux mots. Ceci rend alors les modèles du langage thématiques **adaptatifs**.

Cet article propose d'utiliser deux modèles. Le premier est composé d'unigrammes thématiques basés sur tous les mots du vocabulaire. Le second modèle repose sur la comparaison entre le contenu d'une mémoire cache et les distributions statiques des mots clés des unigrammes thématiques (avec une valeur constante fixée pour chaque thème et qui est assignée à tous les autres mots qui n'entrent pas dans le cache). Une combinaison linéaire des deux modèles peut être utilisée avec différentes approximations de l'historique des mots :

$$\begin{aligned} P(T_j|h_i) &= \alpha_1 P_1(T_j|W_1^{i-1}) \\ &+\alpha_2 P_2(T_j|d(R_j, cc(i-1))) \end{aligned} \qquad (3)$$

où W_1^{i-1} est la séquence $\{w_1, ...w_{i-1}\}$ des premiers $(i-1)$ mots d'un document.

Le premier terme de la formule (3) cumule les informations de l'historique, en contenant tous les mots du vocabulaire. Si ce premier modèle est basé sur les unigrammes, alors :

$$P_1(T_j|W_1^{i-1}) = \prod_{t=1}^{i-1} \frac{P(w_t|T_j)P(T_j)}{\sum_{k=1}^{J} P(w_t|T_k)P(T_k)} \qquad (4)$$

Le problème de ces modèles provient de leur fiabilité limitée et du fait que les mots-clés ont la même importance que ceux qui ne le sont pas.

La caractérisation de thèmes peut être améliorée par l'introduction du second terme de l'équation (3). Alors, un ensemble Γ_j de mots clés est sélectionné pour chaque thème. Sa distribution statistique R_j est obtenue à partir d'un corpus d'apprentissage, et elle est comparée avec la distribution du contenu de $cc(i-1)$ de la mémoire cache, quand le mot w_{i-1} est entré. Le résultat est $d(R_j, cc(i-1))$. C'est l'évaluation de la distance de Kullback-Lieber symétrique qui varie dans le temps, selon l'entrée de nouveaux mots. La probabilité P_2 peut être conditionnée en fonction de la distance entre la distribution du contenu du cache et R_j. Par souci de simplicité, un seul cache est utilisé (un cache différent pour chaque partie de discours pourrait être envisagé).

On suppose que les thèmes forment une partition de l'ensemble des segments. Soit L la taille du vocabulaire. Pour chaque thème, il est possible de trier les L mots selon leur fréquence dans le corpus d'apprentissage et de sélectionner les mots les plus fréquents. Cependant, les mots les plus fréquents qui sont communs à tous les thèmes ne sont pas de bons candidats. Ceux-ci ont été identifiés, et placés dans une liste ; ils ne font pas partie des mots clés. Les premiers $C < L$ mots acceptés de chaque thème sont appelés "*mothèmes*" clés. La comparaison du contenu du cache se limite seulement à ces derniers. Il est alors préférable de normaliser les probabilités des mothèmes clés par rapport aux $L-C$ mots les moins probables auxquels on donne une même probabilité égale à un seuil dans l'unigramme.

Il est possible de faire en sorte que P_2 soit indépendante du thème. En ce sens, la contribution du second terme du mélange de (3) diffère du premier. En fait, plutôt que d'être le produit des probabilités n-grams évaluées sur tout le vocabulaire, il est la probabilité d'une distance évaluée avec des distributions limitées aux mots clés des thèmes.

1. Probabilités dépendantes du thème

Soit $P_{jC} = \sum_{i=1}^{C} P_1(w_i|T_j)$, la somme des probabilités statiques des unigrammes dans le thème T_j. Pour le calcul de la distance, les probabilités de ces C mots sont multipliées par une constante γ_j dépendante du thème. Alors, la probabilité seuil des autres $L-C$ mots est :

$$P_{jr}(w) = \frac{1-\gamma_j P_{jC}}{L-C}$$

Après expérimentation, C est fixé à 4 000 pour chacun des thèmes. De meilleurs choix des C mots clés pour chaque thème restent possibles. Pour peu qu'un étiquetage soit disponible, on peut par exemple choisir de ne retenir que les mots de certaines classes syntaxiques.

En supposant qu'il y ait m mots dans le cache, et que ces mots correspondent à G mots différents, la probabilité d'un mot dans le cache est donnée par :

$$P_{cache}(w) = \frac{n(w)+\beta}{\beta L + m} \tag{5}$$

où $n(w_i)$ est le nombre d'occurrences du mot w_i dans le cache, β est un seuil constant et $\sum n(w) = m$. La probabilité des mots absents du cache est :

$$P_{fC}(w) = \frac{\beta}{\beta L + m} \tag{6}$$

Ainsi, pour avoir une probabilité seuil identique pour les mots absents du cache et ceux des modèles thématiques normalisés, il est suffisant de poser :

$$P_{jr}(w) = P_{fC}(w)$$

$$\frac{1-\gamma_j P_{jC}}{L-C} = \frac{\beta}{\beta L + m}$$

Ce qui donne :

$$\gamma_j = \{1 - \beta\frac{L-C}{\beta L+m}\}\frac{1}{P_{jC}} = \frac{1}{P_{jC}}(1-A) \tag{7}$$

La probabilité normalisée pour un mot dans le thème T_j devient :

$$\begin{aligned} P'(w_k|T_j) &= \frac{n_j(w_k)}{\sum_{i=1}^{C} n_j(w_i)}(1-A) \\ &= (1-A)f'(w_k|T_j) \end{aligned} \tag{8}$$

avec $f'(w_k|T_j) = \frac{n_j(w_k)}{\sum_{i=1}^{C} n_j(w_i)}$.

où $n_j(w_i)$ est le compte des mots dans le thème T_j.

2. Évaluation de la distance

L'évaluation de la distance commence quand le cache contient assez de mots. $d_j(n)$ est la distance relative au thème j après avoir lu les n premiers mots. $P_{cache}(n, w_i)$ est $P_{cache}(w_i)$ à l'insertion dans le cache du n-ème mot du document en cours de traitement. Tous les mots qui ne sont des mots clés dans aucun des thèmes ont la probabilité seuil dans la distribution dépendante du thème et dans le cache. Ainsi, leur contribution à la distance de KL est nulle. Tous ces mots peuvent être ignorés dans le calcul du second terme de la formule (3). La distance de KL symétrique peut alors être calculée comme suit :

$$\begin{aligned} d_j(n) = \sum_{i=1}^{C} &(P_{cache}(n, w_i) - Bf'(w_i|T_j)) \\ &log\left(\frac{P_{cache}(n, w_i)}{Bf'(w_i|T_j)}\right) \end{aligned} \tag{9}$$

Un mot qui n'est mot clé pour aucun thème et qui apparaît dans le cache fournirait une contribution égale dans l'évaluation des distances de chacun des thèmes.

Pour cette raison, ces mots n'ont pas besoin d'entrer dans le cache.

Les coefficients α_1 et α_2 de (3) doivent sommer à 1. Ils sont déterminés par l'interpolation proposée en ([KUH90]). Le coefficient α_2 peut être pris plus petit si le cache n'est pas plein. Il faut noter qu'il est possible que le cache ne contienne aucun des mots clés d'un thème. Si une probabilité nulle était affectée au cache pour chacun des mots clés, la distance aurait une valeur infinie. L'introduction d'une probabilité seuil évite cette situation.

La contribution de la distance de KL pour l'évaluation de la probabilité d'un mot peut être de différents types. Considérons d'abord le vecteur D composé des distances entre le cache et chacun des thèmes :

$$D(n) = \lfloor d_1(n), ..., d_j(n), ...d_J(n) \rfloor$$

La probabilité suivante peut être considérée :

$$P_2(T_j|D(n)) = \frac{P\lfloor D(n)|T_j \rfloor P(T_j)}{\sum_{k=1}^{J} P\lfloor D(n)|T_k \rfloor P(T_k)} \qquad (10)$$

Différentes approximations pour l'évaluation des probabilités antérieures sont possibles. Les expériences avec les probabilités des distances n'ont pas encore été réalisées.

Tab. 1: Thèmes et leurs codes

1	Etranger	5	Affaire, Economie
2	Histoire	6	Culture, Arts, Livres, Media
3	Sciences	7	Politique
4	Sports		

3. Résultats

Trois années des articles du journal *"Le Monde"* ont été utilisées, ce qui représente un total de 60 millions de mots pour un vocabulaire de 500 000 mots. Les thèmes de ces articles ne sont pas connus, mais, grossièrement, les secteurs de rédactions du journal ont été pris comme thèmes. Sept d'entre-eux en ressortent ; leurs intitulés et codes sont définis dans la table 1. Un corpus de test de 1 021 paragraphes pris parmi les articles a été extrait, et n'a pas été utilisé dans l'apprentissage.

Des résultats préliminaires sont présentés dans les tables 2 et 3. U représente le thème retenu par l'unigramme seulement (la probabilité P_1 de la formule (4)), C représente le thème retenu par le modèle cache (dans notre cas c'est une simple normalisation de la distance de KL), S représente le thème d'origine du paragraphe dans le journal, C out représente le cas le cache contient moins de 5 mots clés. Les probabilités sont évaluées mot à mot et les résultats présents sont basés sur les valeurs au dernier mot de chaque paragraphe. Le symbole = représente l'accord, $\neq$ représente le désaccord,

La table 3 résume tous les résultats de l'identification thématique du jeu de test en fonction des différentes stratégies utilisées. La première ligne réfère au cas où le cache et l'unigramme sont en accord avec le label. Dans la deuxième ligne, on ajoute le cas où l'unigramme est en accord avec le label et le cache ne contient pas assez de données pour être pris en compte. La troisième ligne intègre à la deuxième, la décision prise par l'unigramme seulement. La quatrième ligne intègre à la deuxième, la décision prise par le cache seulement, excepté quand le cache ne contient pas assez de mots. La dernière ligne correspond à la décision alterne entre le cache et l'unigramme. La stratégie de combinaison consiste à utiliser le cache pour décision finale seulement s'il contient assez de mots et si la différence entre les scores du cache des premiers et seconds candidats est supérieure à un seuil. Ce dernier est déterminé durant l'apprentissage. Ces résultats préliminaires avec une règle grossière de décision montrent une augmentation des performances grâce à l'utilisation du cache. Avec la contribution du cache, une précision supérieure à 78 % est obtenue.

Conclusion

Les expériences décrites dans cet article montrent les avantages de l'utilisation conjointe d'un modèle cache et d'un modèle unigramme. Leur complémentarité permet une astucieuse combinaison et peut aboutir à une amélioration substantielle de la classification thématique. La tâche de la classification thématique est très délicate à traiter Elle repose en effet sur la décision humaine et son caractère subjectif fait qu'elle est souvent remise en cause par d'autres personnes. Les recherches futures s'orientent vers plusieurs directions. L'emploi d'un lemmatiseur servira à pallier les lacunes du cache en réduisant de fait la taille du vocabulaire. Différents types de probabilités pour les distances peuvent être étudiés étant donné leur utilisation dans les ML adaptatifs pour la reconnaissance automatique de la parole (ASR). L'utilisation de plusieurs caches peut être introduite. Les mots pourront être répartis selon leur classe sémantique connue grâce à l'utilisation d'un programme d'étiquetage sémantique. D'autres utilisations des mots et de leurs classes sémantiques peuvent aussi être envisagées.

Références

[CAR96] B. A. Carlson. Unsupervised topic clustering of switchboard speech messages. Proc. of the *IEEE International Conference on Acoustics*, pages 315–319, Atlanta GA, 1996.

[IMA97] T. Imai R. Schwartz F. Kubala and L. Nguyen. Improved Topic Discrimination of Broadcast News Using a Model of Multiple Simultaneous Topics. Proc. of the *IEEE International Conference on Acoustics*, Speech and Signal Processing, pages 727–730. Munich, Germany, 1997.

[KUH90] R. Kuhn R. De Mori. A cache-based na-

Tab. 2: Résultats préliminaires de classification thématique, en nombre de segments

Thème	U ≠ S C ≠ S U ≠ C	U ≠ S C ≠ S U = C	U = S C ≠ S U ≠ C	U ≠ S C = S U ≠ C	U = S C = S U = C	U ≠ S C out	U = S C out	Total
1	4	10	29	13	75	2	6	139
2	0	0	1	0	6	0	0	7
3	19	4	2	17	48	1	0	91
4	15	7	12	57	21	0	2	114
5	2	3	16	6	129	0	3	159
6	15	21	55	24	244	6	16	381
7	15	21	18	18	51	3	4	130
Total	70	66	133	135	*574*	12	31	1021

Tab. 3: Résultats avec stratégie de combinaison

Stratégie	N	%
U = C = S	574	56,22
+ C out	605	59,26
Deux premiers + Unigramme	738	72,28
Deux premiers + Cache	740	72,48
Stratégie de combinaison	800	78,35

tural language model for speech recognition. In *IEEE Trans. Pattern anal. Machine Intell*, PAMI-12(6), pages 570–582, 1990.

[LIY97] H. Li and K. Yamamishi. Document classification using a finite mixture model. Proc. of the *Conference of the Association for Computational Linguistics*, pages 39–47, Madrid, Spain, 1997.

[PES96] B. Peskin S. Conolly L. Gillick S. Lowe D. McAllaster V. Nagesha P. Van Mulbregt S. Wegmann Improvement in SWITCHBOARD recognition and topic identification. Proc. of the *IEEE International Conference on Acoustic*, Speech and Signal Processing, pages 303–306, Atlanta GA, 1996.

[SPR97] T. Spriet M. El-Bèze. Introduction of Rules into a Stochastic Approach for Language Modeling In NATO ASI series F, Edited by K. M. Ponting, Sprnger Verlag, Berlin New-York, 1997.

[WRI96] J. H. Wright M. J .Carry and E. S. Parris Statistical models for topic identification using phoneme substrings Proc. of the *IEEE International Conference On Acoustics*, Speech and Signal Processing, pages 307–310, Atlanta GA, 1996.