Identification thématique hiérarchique : Application aux forums de discussions

Brigitte Bigi, Kamel Smaïli

Identification thématique hiérarchique : Application aux forums de discussions

Brigitte Bigi, Kamel Smaïli
LORIA - Université Henri Poincaré
Campus Scientifique, BP 239, 54506 Vandœuvre-lès-Nancy
{bigi, smaili}@loria.fr

Mots-clefs – Keywords

Identification thématique, modèles de langage, unigrammes
Topic identification, language modeling, unigrams

Résumé - Abstract

Les modèles statistiques du langage ont pour but de donner une représentation statistique de la langue mais souffrent de nombreuses imperfections. Des travaux récents ont montré que ces modèles peuvent être améliorés s'ils peuvent bénéficier de la connaissance du thème traité, afin de s'y adapter. Le thème du document est alors obtenu par un mécanisme d'identification thématique, mais les thèmes ainsi traités sont souvent de granularité différente, c'est pourquoi il nous semble opportun qu'ils soient organisés dans une hiérarchie. Cette structuration des thèmes implique la mise en place de techniques spécifiques d'identification thématique. Cet article propose un modèle statistique à base d'unigrammes pour identifier automatiquement le thème d'un document parmi une arborescence prédéfinie de thèmes possibles. Nous présentons également un critère qui permet au modèle de donner un degré de fiabilité à la décision prise. L'ensemble des expérimentations a été réalisé sur des données extraites du groupe 'fr' des forums de discussion.

Statistical language modeling attempts to capture the regularities of natural language. The most accurate natural language processing systems still suffer from several shortcomings due to the complexity of natural language and from the weakness of the current language models. It is commonly conjectured that they should benefit from topic adaptation. The topic of the document is then obtained by a topic identification mechanism, but topics thus treated are often of different granularity. This is the reason why it seems appropriate to organize them in a hierarchy. This topic organization implies a development of specific techniques for topic identification. This paper proposes a statistical model based on unigrams to automatically identify the topic of a document among a tree structure of possible topics. We also present a criterion which reflects the degree of reliability of the decision. Experiments were carried out on data extracted from the French newsgroup 'fr'.

1 Introduction

La modélisation statistique du langage est une partie cruciale d'une grande variété d'applications relatives aux technologies du langage, telles que la reconnaissance automatique de la parole (RAP) ou la recherche documentaire. Le but des modèles de langage est de capturer les régularités du langage naturel en estimant les fréquences des mots ou des expressions dans un historique. Ces systèmes souffrent toujours d'imperfections dues à la complexité du langage naturel et à la faiblesse des modèles de langage actuels. Des travaux récents montrent que les modèles de langage doivent tirer bénéfice de la connaissance du thème traité afin de s'y adapter. C'est pourquoi, cet article s'intéresse au problème de l'identification thématique, afin que le thème des documents traités soit déterminé automatiquement.

L'identification thématique a ainsi pour but d'assigner un label thématique à un texte parmi un ensemble de labels possibles. Cet article présente une approche d'identification thématique dans laquelle on exploite les relations sémantiques entre thèmes, par le biais d'une arborescence. Ainsi, par exemple, il peut s'avérer intéressant de spécifier que *football*, *escrime*, *natation* ou encore *plongée* sont des sous-thèmes de *sport*. Un grand nombre de mots sont très probables de façon relativement commune à chacune de ces sous-catégories (comme *tournoi, rencontre, exploit, arbitre*...), ce qui permettra de favoriser les thèmes issus de sport, tandis que certains mots sont spécifiques à l'une ou l'autre des sous-catégories (*but, épée, maillot, tuba*).

Le fait de savoir que les thèmes sont tous relatifs au sport permet à la fois d'éliminer un certains nombre de catégories "concurrentes" pour l'identification du thème mais surtout d'éviter quelques ambiguïtés sur les homonymes (le *tuba* de plongée, ou le *tuba* instrument de musique). Nous pensons que les relations sémantiques établies par la hiérarchie peuvent devenir un atout précieux afin d'obtenir une identification thématique plus fiable. Par ailleurs, lors de la phase d'adaptation, les relations établies entre les modèles de langages thématiques de l'arborescence permettent de compléter certains modèles trop pauvres, comme cela est fait dans (Seymore & Rosenfeld, 1997).

Dans cet article, nous proposons l'utilisation de modèles de langage de type unigrammes, car ces derniers ont déjà montré leur potentiel à discriminer les thèmes dans des applications d'identification thématiques (Li & Yamanishi, 1997; Bigi *et al.*, 2000). Nous proposons des unigrammes thématiques hiérarchiques qui ont pour particularité le fait que les relations entre frères sont favorisées dans le modèle, par l'attribution d'un vocabulaire commun. Par ailleurs, ce modèle utilisera son pouvoir discriminant afin d'auto-évaluer sa décision thématique, en fournissant un degré de fiabilité sur le thème qu'il choisit. Le modèle a été validé sur des données des forums de discussion français. Dans une première section, cet article présente le problème de l'identification thématique et la manière dont il a été décrit dans la littérature. Cette première partie abordre également la notion de hiérarchisation des thèmes. Dans la deuxième section, on présente le modèle d'identification thématique utilisé : un modèle unigramme thématique hiérarchique. Enfin, la dernière section donne les performances concluantes que nous avons obtenues avec notre approche.

2 Positionnement du problème

Les travaux récents (Kneser & Peters, 1997; Martin *et al.*, 1997; Mahajan *et al.*, 1999; Khudanpur & Wu, 1999; Bigi *et al.*, 2000) ont montré que l'adaptation des modèles de langage au

thème du discours permettent une amélioration significative de la perplexité[1] (Jelinek, 1990). Souvent, ce sont des techniques issues du domaine de la recherche documentaire qui sont utilisées. Le problème que nous soulevons dans ces travaux concerne le fait que le thème est obtenu après une analyse de la totalité du document, ce qui n'est évidemment pas envisageable pour une adaptation thématique dynamique dans un système de RAP. Nos travaux se focalisent donc sur des méthodes statistiques qui auront pour motivation de déterminer au mieux le thème d'un document avec le minimum d'information possible. Dans ce cas précis, nos travaux précédents nous ont permis d'observer que les classifieurs bayésiens de type unigrammes obtiennent les meilleures performances d'identification thématique, par rapport à des méthodes classiques.

2.1 Représentation des thèmes dans une hiérarchie

Dans le cas de l'identification thématique hiérarchique, le but est d'exploiter la représentation arborescente des thèmes. La littérature dans le domaine de la hiérarchisation des thèmes reste assez pauvre. L'article principal concernant la reconnaisance automatique de la parole est celui de (Seymore & Rosenfeld, 1997). Dans un premier temps, ils construisent une arborescence de clusters par un système simple à base de mots-clés, puis ils apprennent le modèle qui correspond à chaque cluster. Pour déterminer le "thème" du texte, ils utilisent le classifieur TFIDF (Salton, 1991). Dans le domaine de la RAP, on trouve aussi l'article (Galescu & Allen, 2000) où les auteurs proposent un un "modèle de langage statistique hybride hiérarchique". Dans les deux cas, la combinaison du modèle classique avec le modèle issu de la hiérarchie améliore significativement la perplexité.

On retrouve des hiérarchies de thèmes également dans le domaine de la recherche documentaire sur l'internet, où les documents sont souvent hiérarchisés (Yahoo par exemple). (McCallum *et al.*, 1998) prouvent que les classifieurs bayésiens donnent de meilleures classifications si les données sont hiérarchisées en thèmes. Ce résultat s'appuie sur une expériementation avec des données de certaines sous-catégories du web et des newsgroups.

Les travaux présentés dans ce document concernent le problème de l'identification thématique hiérarchique, dont le but est de déterminer automatiquement le thème d'un texte parmi un ensemble hiérarchisé de thèmes. Ces travaux s'inscrivent dans le cadre d'une amélioration des modèles de langage pour la RAP (figure 1). L'objectif est de résoudre les problèmes de différences de granularité entre les thèmes et de profiter des liens sémantiques entre les thèmes pour l'adaptation des modèles de langage. Malheureusement, les corpus traditionnellement utilisés en RAP ne sont pas hiérarchisés. Cette étude portera sur le groupe "fr" de UseNet.

[1] En général, les modèles de langage pour la reconnaissance de la parole sont évalués en fonction de leur impact sur la précision de la reconnaissance. Néanmoins, ils peuvent être évalués séparément si l'on considère, par exemple, leur capacité de prédiction des mots d'un texte. La mesure la plus utilisée est la *perplexité*. La perplexité d'un modèle de langage dérivée d'un corpus est définie comme suit :

$$PP = 2^{LP(W_1^n)} = P(W_1^n)^{-\frac{1}{n}} \quad (1)$$

où $LP(W_1^n)$ est le logarithme de la probabilité de la séquence de n mots W_1^n attribuée par un modèle de langage bigramme. Pour l'évaluation d'un modèle de langage, on estime les probabilités du modèle avec un ensemble d'apprentissage et on évalue la perplexité avec ce modèle sur un corpus de texte entièrement différent du corpus d'apprentissage.

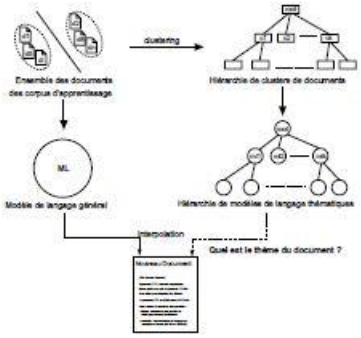

Figure 1: Processus d'intégration d'une hiérarchie de thèmes pour la RAP

2.2 Application : spécificité des forums de discussion

Les "News" sont des forums de discussion fédérés par thème, où, pendant une durée de temps donnée, tous les courriers envoyés sont conservés. Chaque forum est appelé en anglais newsgroup, chaque article d'un newsgroup est appelé une News. Les groupes sont subdivisés en sous-groupes, etc ce qui crée une arborescence dans les groupes. Dans le cadre de l'identification thématique hiérarchique, les news constituent une plate-forme intéressante. En effet, on peut considérer que les données des news se rapprochent des données de parole selon plusieurs aspects. En premier lieu, on peut citer le niveau de langue, qui est souvent familier. Par ailleurs, les messages sont souvent très courts, comprennent de nombreuses abréviations, des fautes d'orthographes et de grammaire, ainsi que des fichiers attachés de toutes natures, autant de difficultés qui peuvent être assimilées aux fautes que commet un système de RAP.

3 Modèles unigrammes pour la classification hiérarchique

3.1 Modèle unigramme thématique classique

On note $W_1^N = \{w_1, w_2, \cdots, w_N\}$ le message dont le thème est à déterminer, parmi les J thèmes prédéfinis. Le but de ce modèle est de déterminer $P(T_j \mid W_1^N)$, c'est-à-dire la probabilité à attribuer à chaque thème en fonction du message. Selon la règle de Bayes, on évalue cette probabilité telle que :

$$P(T_j \mid W_1^N) = \frac{P(T_j).P(W_1^N \mid T_j)}{P(W_1^N)} \tag{2}$$

où $P(T_j)$ est la probabilité *a priori* du thème T_j, et $P(W_1^N \mid T_j)$ représente la probabilité de la séquence de mots W_1^N, étant donné un thème T_j. Celle-ci s'estime comme suit :

$$P(W_1^N \mid T_j) = \prod_{n=1}^{N} P(w_n \mid T_j)$$

inconnu, et correspond au deuxième niveau de repli. L'estimation de ε, ω_{bk} et γ_{jbk} reste un point important du modèle, il est donc important d'en clarifier les différents aspects.

Le modèle doit respecter la contrainte suivante : $\sum_{w_i \in V'_{bk}} P(w_i \mid T_{jbk}) = 1$, où $\{V'_{bk} = V_{bk} + UNK\}$. Ce qui implique :

$$\sum_{w_i \in V_{jbk}} \gamma_{jbk} f(w_i \mid T_{jbk}) + \sum_{w_i \in V_{bk}; w_i \notin V_{jbk}} \omega_{bk} + \varepsilon = 1$$

Comme il n'y a pas de mots inconnus lors de la phase d'apprentissage, $\sum_{w_i \in V_{jbk}} f(w_i \mid T_{jbk}) = 1$, et donc, par conséquent :

$$\gamma_{jbk} = 1 - \varepsilon - \sum_{w_i \in V_{bk}; w_i \notin V_{jbk}} \omega_{bk} \tag{5}$$

En pratique, on attribuera à ω_{bk} la valeur d'un ratio par rapport à la plus petite des probabilités observées parmi les frères. Ceci signifie que pour chacun des mots qui n'appartiennent pas à un thème donné mais à un des thèmes frères, une distribution de probabilité uniforme sur les thèmes frères sera attribuée. La valeur ε, quant à elle, devra prendre une probabilité très petite (plus petite que le plus petit des ω_{bk}). On peut donc résumer la particularité de ce modèle en énonçant que la valeur ε est la même pour tous les mots inconnus de l'arbre, alors que ω_{bk} dépend du groupe de frères.

3.2.2 Attribution d'un thème au nouveau document

La probabilité du thème T_{jbk} étant donné le document W_1^N est evaluée de la même manière que dans un unigramme classique, où les thèmes sont remplacés par les nœuds de l'arbre :

$$P(T_{jbk} \mid W_1^N) = \frac{P(T_{jbk}).P(W_1^N \mid T_{jbk})}{P(W_1^N)} \tag{6}$$

où $P(T_{jbk})$ est la probabilité *a priori* du thème T_{jbk}.

De même $P(W_1^N \mid T_{jbk}) = \prod_{n=1}^{N} P(w_n \mid T_{jbk})$ est la probabilité de la séquence de mots $W_1^N = w_1, \ldots, w_n$, évaluée comme le produit des probabilités de chaque mot dans le thème T_{jbk}. On obtient donc une distribution des probabilités thématiques de l'arbre, étant donné un message observé.

3.2.3 Auto-évaluation du modèle

Notre objectif est de proposer le thème qui correspond au mieux au message ennoncé. Un apport supplémentaire qui peut s'avérer un atout intéressant, est de faire en sorte que le modèle associe un degré de fiabilité à la décision thématique qu'il prend. Nous proposons que le modèle associe l'un des critères suivants :

1. la décision est certaine,
2. la décision est médiane,
3. la décision est incertaine.

Le problème est donc d'obtenir $P(w \mid T_j)$, la probabilité d'un mot w dans un thème T_j, pour chaque mot w du vocabulaire V_j de T_j :

$$P(w \mid T_j) = \begin{cases} \gamma_j f(w \mid T_j) & si\ (w \in V_j) \\ \varepsilon & sinon \end{cases} \quad (3)$$

où $\sum_{w_i \in V_j} \gamma_j f(w \mid T_j) + \varepsilon = 1$, $f(w \mid T_j)$ est la fréquence du mot w dans T_j, apprise sur un corpus dédié au thème T_j, et, γ_j est un cœfficient de normalisation qui assure que la distribution des probabilités $P(w \mid T_j)$ somme à 1. Comme il n'y a pas de mot inconnu lors de l'apprentissage, la valeur de ε sera donc attribuée à tous les mots n'appartenant pas au vocabulaire de T_j lors de la phase de test. Elle représente la probabilité du mot inconnu, noté UNK.

3.2 Modèles unigrammes thématiques hiérarchiques

La hiérarchie des thèmes est présentée sous la forme d'un arbre (figure 2). A chaque niveau k, on définit des groupes de nœuds frères, notés b (les frères, étant, par définition, des nœuds de même père). Notons également T_{jbk}, le j-ème thème, du b-ème groupe de frères, du k-ème niveau, ainsi que T_{bk} un groupe de frères. On notera également V_{jbk} le vocabulaire du thème T_{jbk}, déduit du corpus d'apprentissage de ce thème.

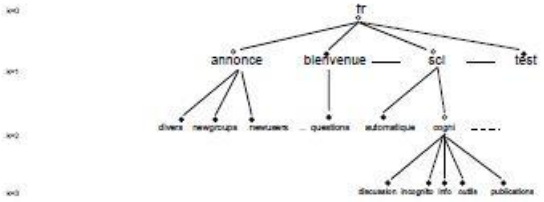

Figure 2: Exemple d'une hiérarchie de thèmes, groupe 'fr' de UseNet

3.2.1 Distribution de probabilités des nœuds

Pour rendre compte des relations sémantiques fraternelles, nous proposons que chaque thème d'un groupe de frères donné soit représenté par le même vocabulaire, issu de l'union des vocabulaires de chacun des thèmes frères. Ce vocabulaire est noté V_{bk}. Ceci implique que certains mots n'auront pas été observés durant l'apprentissage et par conséquent $f(w \mid T_j) = 0$. Le modèle classique des unigrammes de l'équation (3) est donc modifié en introduisant un deuxième niveau de repli (back-off), tel que :

$$P(w \mid T_{jbk}) = \begin{cases} \gamma_{jbk} f(w \mid T_{jbk}) & si\ (w \in V_{jbk}) \\ \omega_{bk} & sinon\ si\ (w \in V_{bk}) \\ \varepsilon & sinon \end{cases} \quad (4)$$

où γ_{jbk} est un cœfficient de normalisation. ω_{bk} représente le premier niveau de repli, qui prend en compte tous les mots du vocabulaire ayant une fréquence nulle. ε est la probabilité du mot

Cette auto-évaluation du modèle est effectuée avec le pouvoir discriminant relatif à sa décision. Celui-ci est évalué par l'écart entre la probabilité attribuée au meilleur thème et la probabilité attribuée au deuxième meilleur :

$$\Delta = P_b(T_{jbk} \mid W_1^N) - P_{2b}(T_{jbk} \mid W_1^N)$$

Cet écart est ensuite comparé à deux seuils δ_1 et δ_2 fixés empiriquement afin de déterminer la fiabilité, selon l'algorithme suivant :

si ($\Delta > \delta_1$)
alors
    décision certaine
sinon
    si ($\Delta > \delta_2$)
    alors
        décision médiane
    sinon
        décision incertaine
    finsi
finsi

4 Expérimentations

4.1 Les données

Nous avons choisi, pour cette étude, de nous restreindre aux groupes de langue française, placés dans une hiérarchie dont la racine est "fr.". Nos données couvrent une période de plusieurs mois chevauchant les années 2000 et 2001. Ce corpus représente un total de 2 Go, avec plus d'1 million d'articles. Il est composé de 365 newsgroups, dont 307 feuilles dans lesquelles on cherchera à poster l'article. Le problème principal relatif à ces données concerne le fait que les articles postés dans les news sont entachés d'erreurs. Cette contrainte implique l'application d'un ensemble de pré-traitements afin d'extraire, autant que possible, les données intéressantes de chacun des articles, c'est-à-dire, le corps du texte sans les "bruits" associés. Parmi ces pré-traitements, il y a notamment une phase de segmentation en mots réalisée en comparaison à un lexique, et certains mots sont regroupés dans des classes selon deux possibilités :

- en référence à un lexique spécialisé[2], pour les noms personnels, les ponctuations, les villes, les pays et les "smileys" ;
- en utilisant des informations syntaxiques dans le cas des adresses électroniques, des adresses internet, des heures, des prix et des nombres (obtenus à partir d'un ensemble de règles).

[2]Les lexiques ont été constitués semi-manuellement à partir de données collectées sur internet.

4.2 Résultats

Etant donné un article, notre but est de proposer le (ou les) forum de discussion qui semble le plus adéquat, parmi l'ensemble des feuilles de l'arbre. Afin d'évaluer notre modèle, nous avons comparé le (ou les) nœud ainsi proposé à ceux dans lesquels les articles du corpus de test ont été envoyés. Les résultats sont donc donnés sous forme de rappel et précision tels que :

rappel : Ratio entre le nombre de thèmes détectés correctement et le nombre de thèmes à détecter ;

précision : Ratio entre le nombre de thèmes détectés correctement et le nombre de thèmes détectés.

4.2.1 Première expérimentation

Le tableau 3 présente les résultats obtenus dans ce cadre d'identification thématique hiérarchique. Dans ce tableau, on peut observer que le modèle unigramme hiérarchique augmente les performances de 2 % de rappel et de précision par rapport à un unigramme classique. Ce résultat fait mention uniquement de la prise en compte des relations sémantiques entre frères : l'augmentation légère des résultats nous permet de vérifier positivement que nous sommes sur la bonne voie. En particulier, on peut observer les performances obtenues par deux sous-groupes relatifs à linux (niveau $k = 4$), et deux autres relatifs à la biologie (niveau $k = 2$). On observe que les groupes linux obtiennent des taux d'identification thématique très intéressants, alors que les groupes de biologie sont mal identifiés. Ces différences importantes de performances peuvent s'expliquer en observant la taille de leur vocabulaire V_{jbk} mis en rapport avec le nombre de documents du corpus d'apprentissage. Ainsi, on peut observer qu'avec des corpus d'apprentissage de tailles proches, les deux thèmes se trouvent très différemment représentés d'un point de vue de leur vocabulaire, ceci étant dû à leur niveau de spécialisation. Par conséquent, il semble évident que si linux est très bien reconnu, c'est parce qu'il dispose de suffisamment de données pour être statistiquement significatif, contrairement à biologie.

Newsgroup	Rappel	Précision	Nb news apprentissage	Nb news test	V_{jbk}
Unigramme classique, sur :					
- tout le corpus de test	0,33	0,37	+1 Million	59157	425248
Unigramme hiérarchique, sur :					
- tout le corpus de test	0,35	0,39	+1 Million	59157	425248
- fr.comp.os.linux.annonces	0,71	0,71	124	7	7954
- fr.comp.os.linux.configuration	0,94	0,97	16111	931	4536
- fr.bio.pharmacie	0,05	0,06	684	35	10363
- fr.bio.medecine	0,34	0,46	14316	490	44970

Figure 3: Performances d'identification thématique sur quelques thèmes

4.2.2 Seconde expérimentation

Dans cette seconde expérimentation, nous introduisons de nouvelles notions relatives à la valeur de rappel :

- "rappel exact" signifie que le thème du modèle doit être rigoureusement celui de la solution, celui-ci correspond au rappel de l'expérience précédente ;
- "rappel voisin" signifie que le thème du modèle peut être soit exact, soit le frère, soit le père, soit le fils de la solution ;
- "rappel branche" signifie que le thème du modèle est dans la même branche que le thème solution (c-à-d les niveaux $k = 1$ égaux).

	Rappel exact	Rappel voisins	Rappel branche	Précision	Nombre de documents test
Tout le corpus	0,35	0,45	0,60	0,39	59 157 (soit 100 %)
Classe incertaine	0,17	0,27	0,44	0,19	30 970 (soit 52 %)
Classe médiane	0,52	0,62	0,75	0,57	22 328 (soit 38 %)
Classe certaine	0,71	0,79	0,87	0,77	5 859 (soit 10 %)

Figure 4: Performances d'identification thématique de l'unigramme hiérarchique incluant l'auto-évaluation

Les résultats sont présentés à la première ligne du tableau 4. Avec un rappel de 0,60 sur la branche solution, on remarque que le modèle, même s'il ne trouve pas le "bon" nœud, trouve dans la majorité des cas le domaine thématique abordé.
Les lignes suivantes (tableau 4) décomposent ce résultat, lorsque le modèle propose un degré de fiabilité associé à sa décision thématique. Ces résultats sont intéressants, car on voit bien que lorsque le pouvoir discriminant du modèle est important, la solution proposée par le modèle est souvent correcte avec un rappel exact égal à 0,71. Avec un rappel sur la branche de 0,87, on voit également que dans ces cas, même lorque l'on ne prédit pas le thème exact, on est quand même capable d'apporter une solution avoisinante, ou tout au moins d'indiquer la branche à suivre avec un risque d'erreur très faible.

5 Conclusion

Dans cet article, nous avons présenté un modèle unigramme thématique hiérarchique pour l'identification thématique hiérarchique. Ce modèle offre des performances légèrement supérieures à celles obtenues avec un unigramme classique, dû ûau fait que les relations entre frères sont prises en compte à travers une union de leur vocabulaire, et à travers l'insertion d'un facteur de repli à deux niveaux. Nous avons également montré que même si les performances sur le nœud exact ne sont pas élevées, elles augmentent nettement lorque l'on se compare à la branche choisie. Concernant l'ensemble de ces résultats, il est important de rappeler que le thème choisi par le modèle est comparé avec celui dans lequel l'article a été posté par l'expéditeur. Ce critère

n'est pas fiable, car il existe de nombreux groupes de news, et les utilisateurs n'ont pas toujours connaissance de leur existence, ou de leur contenu réel. Ceci implique que la décision relative au groupe dans lequel son article doit être expédié n'est pas fiable, mais ce critère de comparaison est le seul dont nous disposons. Nous avons également pu observer que le pouvoir discriminant du modèle est un critère suffisamment pertinent pour permettre au modèle de s'auto-évaluer et ainsi, de donner non seulement le thème du nouveau document, mais aussi, d'y associer un degré de confiance. Ainsi, dans près de la moitié des articles, on trouve la branche avec un rappel de plus de 0,75, et le nœud exact avec un rappel de plus de 0,52.

Ces résultats sont encourageants. Ils laissent entrevoir, entre-autres, la possibilité de leur utilisation dans un système de reconnaissance automatique de la parole. Dans ce cas, l'auto-évaluation du modèle est un facteur important qui permettra de n'introduire le modèle thématique que dans les cas où le thème est obtenu avec confiance. Différentes voies de recherche restent à explorer pour améliorer encore ces travaux. Notamment, ils pourraient être mis en place sur une arborescence créée automatiquement. Dans ce cas, la méthode d'identification thématique intègrera certains des paramètres qui ont permis de constituer l'arbre, afin que les méthodes de classification et d'identification soient relativement homogènes.

Références

BIGI B., DE MORI R., EL-BÈZE M. & SPRIET T. (2000). A fuzzy decision strategy for topic identification and dynamic selection of language models. *Special Issue on Fuzzy Logic in Signal Processing, Signal Processing Journal*, 80(6), 1085–1097.

GALESCU L. & ALLEN J. (2000). Hierarchical statistical language models: experiments on in-domain adaptation. In *Proceedings of the 6th International Conference on Spoken Language Processing (ICSLP'2000)*, p. 16–20, Beijing, China.

JELINEK F. (1990). Self-organized language modeling for speech recognition. *Readings in Speech Recognition, A. Waibel and K-F. Lee editors*, p. 450–506.

KHUDANPUR S. P. & WU J. (1999). A maximum entropy language model integrating n-gram and topic dependencies for conversational speech recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, volume I, p. 2192, Phoenix, Arizona.

KNESER R. & PETERS J. (1997). Semantic clustering for adaptive language modeling. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, p. 779–783, Munich, Germany.

LI H. & YAMAMISHI K. (1997). Documentation classification using a finite mixture model. In *Conference of the Association for Computational Linguistics*, p. 39–47, Madrid, Spain.

MAHAJAN M., BEEFERMAN D. & HUANG X. D. (1999). Improved topic-dependent language modeling using information retrieval techniques. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, volume I, p. 2391, Phoenix, Arizona.

MARTIN S. C., LIERMANN J. & NEY H. (1997). Adaptive topic-dependent language modeling using word-based varigrams. In *Proceeding of the European Conference On Speech Communication and Technology*.

MCCALLUM A., ROSENFELD R., MITCHELL T. & NG A. (1998). Improving text classification by shrinkage in a hierarchy of classes. In *International Conference on Machine Learning*.

SALTON G. (1991). Developments in automatic text retrieval. *Science*, **253**, 974–980.

SEYMORE K. & ROSENFELD R. (1997). Using story topics for language model adaptation. In *Proceeding of the European Conference On Speech Communication and Technology*.