Nouvelle approche de la sélection de vocabulaire pour la détection de thème

Armelle Brun, Kamel Smaïli, Jean-Paul Haton

TALN 2003, Batz-sur-Mer, 11–14 juin 2003

Nouvelle approche de la sélection de vocabulaire pour la détection de thème

Armelle BRUN, Kamel SMAILI, Jean-Paul HATON
LORIA BP 239 54506 Vandœuvre-Lès-Nancy, France -
Tel : (33|0) 3-83-59-20-97, Fax :(33|0) 3-83-41-30-79
{brun, smaili, jph}@loria.fr

Mots-clefs – Keywords

Détection de thème, création de vocabulaire, combinaison
Topic detection, vocabulary creation, combination

Résumé - Abstract

En reconnaissance de la parole, un des moyens d'améliorer les performances des systèmes est de passer par l'adaptation des modèles de langage. Une étape cruciale de ce processus consiste à détecter le thème du document traité et à adapter ensuite le modèle de langage. Dans cet article, nous proposons une nouvelle approche de création des vocabulaires utilisés pour la détection de thème. Cette dernière est fondée sur le développement de vocabulaires spécifiques et caractéristiques des différents thèmes. Nous montrons que cette approche permet non seulement d'améliorer les performances des méthodes, mais exploite également des vocabulaires de taille réduite. De plus, elle permet d'améliorer de façon très significative les performances de méthodes de détection lorsqu'elles sont combinées.

One way to improve performance of Automatic Speech Recognition (ASR) systems consists in adapting language models. We are particularly interested in adapting language models to the topic related in data. Before adapting the language model, this topic has to be detected. In this work, we present a new way to create vocabularies used to detect the topic in a given text. This new method results in the improvement of topic detection performance of the methods studied, it also results in the reduction of the vocabulary size required. Finally, we show a large improvement of the performance when combining topic identification methods, when new vocabularies are used.

1 Introduction

Les systèmes de Reconnaissance Automatique de la Parole (RAP) actuels atteignent des performances intéressantes dans des applications ciblées. Les données en entrée d'un système de RAP se présentent sous la forme d'un signal acoustique correspondant à une phrase prononcée. Ces données sont tout d'abord traitées par un module de traitement du signal acoustique. Malgré des performances très élevées de ce dernier, son utilisation seule ne permet pas d'obtenir des résultats de reconnaissance suffisamment élevés. En effet, bien que les phrases proposées par le système soient très proches acoustiquement de la suite de mots prononcée, ces dernières sont bien souvent syntaxiquement incorrectes. Pour palier les faiblesses de ce module, un second modèle est exploité, en complément de celui-ci. Il a pour fonction de modéliser la langue et aura donc pour rôle d'affiner les différents scores des phrases proposées par le module acoustique, c'est le modèle de langage.

Les modèles statistiques de langage des systèmes de RAP modélisent la langue sous forme probabiliste. Plus particulièrement, ils évaluent la probabilité d'apparition d'un mot sachant les mots le précédant dans la phrase (son historique). Les modèles les plus utilisés à l'heure actuelle sont les modèles de langage dits $n-$grammes. Ils estiment la probabilité d'apparition d'un mot uniquement en fonction des $n-1$ derniers mots le précédant. Pour des raisons d'estimation de probabilités et de stockage, la taille de l'historique pris en compte ($n-1$) ne dépasse généralement pas 3. Le reproche fait à ce type de modèle est justement de prendre en compte un historique trop restreint. Pour pallier cet inconvénient, de nombreux modèles ont été développés dans le but de mieux prédire les mots.

Une des approches utilisées pour l'amélioration de la qualité de ces modèles consiste à adapter le modèle de langage du système aux caractéristiques du texte en cours de traitement. Dans ce cadre, nous nous intéressons à l'adaptation des modèles de langage au thème traité dans le document. Nous considérons, en effet, que le vocabulaire utilisé dans un texte est dépendant du thème traité dans ce dernier. Dans le cadre de l'adaptation, nous choisirons donc d'exploiter un modèle de langage représentatif du thème traité dans le texte. Par conséquent, l'étape cruciale de cette adaptation est la recherche du thème traité dans le document.

Dans cet article, nous allons tout d'abord présenter le domaine de la détection de thème, et notamment les deux grands paramètres influençant la détection de thème : le vocabulaire et la méthode de détection de thème. Après avoir introduit les données sur lesquelles nous travaillons, nous allons exposer les performances en détection de thème obtenues par les méthodes de l'état de l'art. Ensuite, nous présenterons une nouvelle méthode de sélection de vocabulaire et nous étudierons les performances obtenues avec les nouveaux vocabulaires. Enfin, nous concluerons et nous présenterons quelques perspectives à nos travaux.

2 La détection de thème

Soient un document donné et un ensemble prédéfini de thèmes, la détection de thème a pour but de rechercher le(s) thème(s) traité(s) dans ce document. Cette dernière est fonction de deux paramètres principaux qui sont le vocabulaire utilisé et la méthode de détection de thème. Le vocabulaire définit l'ensemble des éléments caractéristiques d'un thème. Classiquement, c'est sur la base de l'ensemble des mots le constituant que la plupart des méthodes fondent leurs principes de détection.

2.1 Vocabulaire

Dans un texte, toutes les informations présentes ne sont pas utiles pour la détection de thème. Par exemple, il est assez intuitif que dans un document contenant la phrase « le gardien de but n'a pas réussi à arrêter le tir », les deux occurrences du mot « le » n'apportent aucune information quant au thème traité dans un texte. A l'opposé, l'expression « gardien de but » est très importante et suggère que le texte traite de sport.

Dans le cadre de la détection de thème, un vocabulaire recense l'ensemble des caractéristiques des thèmes utiles pour cette tâche. Dans le domaine de la recherche d'informations, (Lewis, 1992) a montré que l'utilisation du mot comme unité de représentation d'un document semble être adaptée pour des tâches de classification. Pour cette raison, la majorité des méthodes de détection de thème utilise le mot comme unité de représentation du document (représentation *bag of words*). Par conséquent, les vocabulaires utilisés seront eux aussi composés de mots, ceux les plus utiles pour la détection de thème. Dans nos travaux, nous considérons les mots sous forme fléchie, des travaux précédents (Frakes & Baeza-Yates, 1992) n'ayant montré aucun gain dans l'emploi de lemmes.

La question qui se pose alors est de savoir quels sont ces mots. Il existe dans la littérature plusieurs méthodes permettant de trouver ces ensembles de mots (qui composeront le vocabulaire). Nous présentons ici quatre méthodes de sélection de vocabulaire, parmi les plus étudiées.

2.1.1 Fréquence des mots

Dans le cas d'un vocabulaire sélectionné par fréquence de mots, on calcule, pour chacun des mots du corpus d'apprentissage, sa fréquence d'apparition. Le vocabulaire sera ensuite composé des mots de fréquence élevée. On considère dans ce cas que plus les mots sont fréquents à l'apprentissage, plus ils sont utiles pour la détection de thème.

2.1.2 Fréquence de document des mots

Dans ce cas, on ne prend pas en compte la fréquence des mots à l'apprentissage, mais le nombre de documents dans lesquels chaque mot est apparu. Le vocabulaire résultant sera composé des mots apparus dans le plus grand nombre de documents.

2.1.3 Information mutuelle

La mesure d'information mutuelle quantifie le lien existant entre un mot et un thème. Plus précisément, elle évalue l'influence qu'a, sur le thème d'un texte, la présence d'un mot dans ce texte. Pour un mot et un thème donnés, elle est évaluée de la façon suivante (Seymore *et al.*, 1998) :

$$I(w_i, T_j) = \log P(w_i \mid T_j) - \log P(w_i) \tag{1}$$

Classiquement, pour un mot donné, on calcule sa valeur d'information mutuelle avec chacun des thèmes. Ensuite, ces valeurs sont combinées afin d'obtenir une valeur unique pour chaque mot. (Yang & Pedersen, 1997) montre que dans ce cas la meilleure façon de combiner consiste

à retenir, pour chaque mot, la valeur d'information mutuelle maximale parmi l'ensemble des thèmes.

2.1.4 Gain d'information

Le gain d'information (également appelé information mutuelle moyenne) (Mitchell, 1996), permet, tout comme la mesure d'information mutuelle, de quantifier le lien existant entre un mot et un thème mais ne prend pas seulement en compte l'influence qu'a l'apparition d'un mot sur un thème, il considère également sa non apparition, etc. La mesure de gain d'information se calcule de la façon suivante :

$$IG(w_i, T_j) = \sum_{T \in \{T_j, \bar{T}_j\}} \sum_{w \in \{w_i, \bar{w}_i\}} P(w, T) \log \frac{P(w, T)}{P(w)P(T)} \quad (2)$$

Comme dans le cas de l'information mutuelle, pour un mot donné, on a une valeur par thème traité. Dans ce cas, (Yang & Pedersen, 1997) montre qu'il faut utiliser la moyenne pondérée des valeurs de gain d'information entre le mot et chaque thème.

Pour l'ensemble des quatre mesures présentées ici, la qualité de chacun des mots dans le langage est calculée. Celle-ci ne tient pas compte des caractéristiques des mots dans les thèmes, elle est évaluée tous thèmes confondus. Le vocabulaire résultant sera composé de l'ensemble des mots ayant, selon la mesure choisie, les valeurs les plus élevées.

2.2 Méthodes de détection de thème

Le second paramètre dans cette approche est la méthode de détection de thème, qui définit la façon dont les informations (mots) présentes dans les textes (suivant le vocabulaire utilisé) sont exploitées.

Nous avons décidé d'étudier un ensemble de méthodes de l'état de l'art parmi les plus anciennes (TFIDF), les plus performantes (cache et unigramme) ainsi que les plus récentes (SVM). Nous avons également étudié une méthode issue du domaine de la RAP (perplexité). Toutes ces méthodes ont été largement présentées dans la littérature, pour cette raison nous ne nous attarderons pas sur leur présentation.

Sachant un vocabulaire donné, chacun des thèmes est tout d'abord schématisé sous la forme d'un vecteur où chaque élément représente la fréquence d'un mot du vocabulaire dans le corpus d'apprentissage du thème. De la même façon, le document de test (celui dont on recherche le thème) est représenté sous forme de vecteur. L'ensemble des méthodes présentées exploite ces représentations vectorielles.

2.2.1 Le classifieur TFIDF

Le classifieur TFIDF (Salton, 1991) est la référence dans le domaine, celui-ci étant un des premiers modèles à avoir été développé. Dans le cas du classifieur TFIDF, chacun des éléments des vecteurs est pondéré par un facteur reflétant la proportion de thèmes dans lequel le mot est présent. Ensuite, une distance cosine (3) est calculée entre le vecteur représentant le document

et celui de chacun des thèmes. Le thème correspondant à la distance la plus faible sera celui affecté au document.

$$sim(D_j, D_i) = \frac{\sum_{k=1}^{|V|} d_{jk} d_{ik}}{\sqrt{\sum_{k=1}^{|V|} (d_{jk})^2 \sum_{k=1}^{|V|} (d_{ik})^2}} \quad (3)$$

2.2.2 Le modèle unigramme

Dans le modèle unigramme (McDonough *et al.*, 1994), une distribution de probabilités des mots est calculée pour chaque thème. Ensuite, la probabilité de chaque thème est calculée (4), le thème correspondant à la probabilité *a posteriori* la plus élevée sera le thème retenu.

$$P(T_j \mid W_1^N) = \frac{P(T_j) P(W_1^N \mid T_j)}{\sum_{k=1}^{J} P(T_k) P(W_1^N \mid T_k)} \quad (4)$$

2.2.3 Le modèle cache

Le modèle cache (Bigi *et al.*, 2000), dérive lui aussi une distribution de probabilités des mots dans chacun des thèmes, mais également des mots dans le document de test (plus précisément d'une fenêtre cache des mots du test). Ensuite, la distance de Kullback-Leibler symétrique est calculée entre la distribution des mots dans le document de test et celle des mots dans les thèmes. Le thème retenu sera celui correspondant à la distance la plus faible. Pour de plus amples détails sur ce modèle, voir (Bigi *et al.*, 2000).

2.2.4 Les Machines à Vecteur Support (SVM)

Contrairement aux trois autres méthodes déjà présentées, la méthode SVM (Vapnik, 1995) traite le cas biclasse. Dans ce cas, elle oppose le thème en cours de traitement à l'ensemble des autres thèmes. Sachant une représentation dans un espace donné, des documents du thème ainsi que de l'ensemble des autres documents, on recherche l'hyperplan optimal séparant les deux ensembles de données. L'originalité des SVM est qu'elles cherchent à maîtriser l'erreur en généralisation. Pour traiter le cas où plus de deux classes (thèmes) sont utilisées, une étape de recombinaison des scores est ensuite nécessaire pour retrouver le thème d'un document donné.

2.2.5 La perplexité

La perplexité (Jelinek & Mercer, 1980) est issue du domaine de la reconnaissance de la parole. La mesure de perplexité permet de mesurer l'adéquation entre un modèle de langage et un document donné. Si l'on développe un modèle de langage par thème et que l'on calcule la valeur de perplexité pour chacun des modèles de langage de thème sur le document de test, alors le thème correspondant à la perplexité minimale sera considéré comme étant celui du thème.

3 Résultats

3.1 Données

Les données sur lesquelles nous travaillons sont extraites de 5 années (1987-1991) du journal *Le Monde* et sont divisées en 7 thèmes. Des ces dernières (environ $86M$ mots) nous avons extrait les données de test, le reste formant le corpus d'apprentissage. Dans nos travaux, nous nous situons dans le cadre de la recherche d'un seul thème dans un document. Cependant, les données sont présentées sous la forme d'articles de journaux et nombre d'articles journalistiques traitent de plusieurs thèmes. Nous faisons l'hypothèse qu'il est fort probable qu'un paragraphe donné ne traite que d'un thème. Par conséquent, nous considérons que le corpus est constitué d'une suite de paragraphes et que l'identification porte sur chaque paragraphe et non sur l'article entier. Le corpus de test regroupe un peu plus de 800 paragraphes, tirés aléatoirement des données, étiquetés à la main par une unique personne. Les figures 1 et 2 présentent les proportions, en fonction des thèmes, des données d'apprentissage et de test.

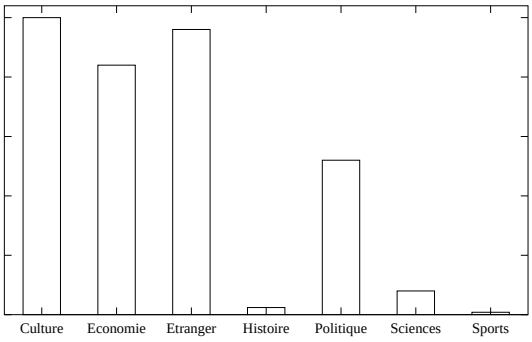

FIG. 1 – Répartition des données d'apprentissage en fonction des thèmes

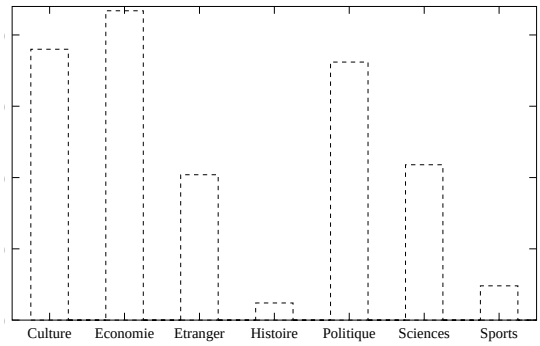

FIG. 2 – Répartition des données de test en fonction des thèmes

3.2 Performances

Les performances en détection de thème sont évaluées par rapport au taux de paragraphes dont l'étiquette a été retrouvée par le module de détection de thème. Pour une méthode de détection de thème et de sélection de vocabulaire fixées, nous évaluons les performances en détection de thème. Un paramètre supplémentaire intervient, celui correspondant à la taille du vocabulaire choisi, celle-ci ayant une grande influence sur les performances. La figure 3 présente, sachant une méthode de détection de thème (unigramme) et une méthode de sélection de vocabulaire (information mutuelle) fixées, l'évolution des performances en fonction de la taille du vocabulaire. Dans ce cas, la taille optimale du vocabulaire est de $10K$ mots. Si le nombre de mots retenus est inférieur ou supérieur, les performances sont plus faibles.

Par conséquent, pour chacune des méthodes de détection de thème, chacune des méthodes de sélection de vocabulaire et un ensemble de tailles de vocabulaire, nous avons évalué les performances en détection de thème. Dans le tableau 1 nous présentons seulement les performances maximales obtenues par association du meilleur vocabulaire et de chaque méthode. La méthode de sélection de vocabulaire ainsi que la taille correspondantes sont également précisées.

Nous pouvons remarquer que l'ensemble des méthodes atteint des performances supérieures à 74%. De plus, les vocabulaires optimaux sont composés de plus de $30K$ mots. Les deux mé-

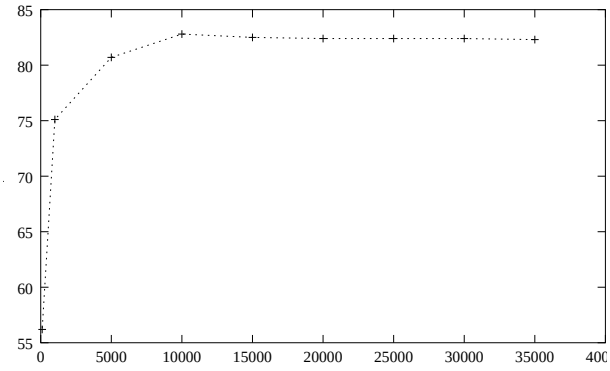

FIG. 3 – Evolution des performances du modèle unigramme utilisant un vocabulaire sélectionné par information mutuelle en fonction de la taille du vocabulaire retenue

Méthode de détection de thème	*Méthode de sélection de vocabulaire*	*Taille de vocabulaire*	*Performances*
Unigramme	Fréquence de document	$30K$	**83.1%**
TFIDF	Fréquence de mots	$30K$	74.3%
Cache	Fréquence de mots	$34K$	**82.5%**
Perplexité	Fréquence de mots	$64K$	79.0%
SVM	Information mutuelle	$40K$	78.3%

TAB. 1 – Performances de chacune des méthodes de détection de thème en fonction de la meilleure méthode de sélection de vocabulaire

thodes les plus performantes sont le modèle cache et le modèle unigramme, ce dernier atteignant des performances de 83.1%.

4 Nouvelle méthode de sélection de vocabulaire

Comme nous l'avons déjà mentionné, dans les méthodes classiques de sélection de vocabulaire, la mesure de la qualité d'un mot est définie dans le langage, *i.e.* tous thèmes confondus. Les meilleurs mots sont ensuite retenus.

Nous considérons, de notre côté, que le vocabulaire ne doit pas être défini au niveau de la langue en général mais plutôt au niveau des thèmes. Nous sommes convaincus que chaque thème a un vocabulaire qui lui est propre et que les performances en détection de thème pourraient être améliorées si de tels vocabulaires pouvaient être pris en compte. Prenons par exemple le cas des deux mots « temps » et « match ». Le mot « temps » est un mot très courant dans l'ensemble des thèmes. Si la méthode de sélection de vocabulaire est celle par fréquence de mots, ce mot sera sélectionné pour composer le vocabulaire. Cependant, la présence de celui-ci dans un document de test ne nous permettra pas de déterminer de façon efficace le thème du texte puisqu'il est fréquent dans l'ensemble des thèmes. A l'opposé, le mot « match » a une fréquence faible dans l'ensemble des thèmes sauf pour le thème Sports. Dans le cas classique, ce mot ne sera pas conservé puisqu'il n'est pas assez fréquent dans le corpus d'apprentissage. A l'opposé, si l'on raisonne au niveau du thème, le mot « match » sera retenu pour le vocabulaire du thème Sports, puisque fréquent dans ce dernier.

Méthode de détection de thème	*Méthode de sélection de vocabulaire*	*Taille de vocabulaire*	*Performances*
TFIDF	Information mutuelle	$5K$	**74.4%**
Unigramme	Information mutuelle	$15K$	**83.4%**
SVM	Gain d'information	$22K$	**78.7%**

TAB. 2 – Performances en détection de thème de trois méthodes pour des vocabulaires de thème de taille équivalente

Dans cette optique, nous exploitons les mesures présentées dans la section 2.1. Celles-ci seront définies non plus au niveau du langage mais au niveau du thème. Ainsi, pour un thème donné, nous évaluons la mesure pour l'ensemble des mots de l'apprentissage du thème. Nous conservons ensuite les mots ayant les valeurs les plus élevées. Nous obtenons dans ce cas un vocabulaire pour chacun des thèmes traités. Nous effectuons ensuite l'union de ces derniers afin de former le vocabulaire utilisé pour la représentation des données.

Dans ce cas, la même question que dans le cas d'un vocabulaire défini au niveau global se pose : combien de mots doit-on conserver pour chaque vocabulaire ? Nous présentons maintenant un cas spécifique pour les vocabulaires de thèmes : nous conservons un nombre identique de mots par thème.

4.1 Nombre identique de mots par thème

Dans le cas qui vient d'être présenté, nous créons les vocabulaires de thème et nous retenons le même nombre de mots pour chaque thème. Ensuite, pour chacune des méthodes de détection de thème et pour chaque méthode de sélection de vocabulaire, nous avons étudié les performances en détection de thème en fonction du nombre de mots retenu par thème.

Le tableau 2 présente les meilleures performances associées à trois des méthodes de détection de thème, ainsi que la méthode de sélection de vocabulaire associée et la taille du vocabulaire correspondante.

Nous pouvons remarquer que pour les trois méthodes étudiées, les performances se sont améliorées. Cette amélioration n'est cependant pas statistiquement significative (entre $+0.1\%$ et $+0.4\%$). Un des points importants que nous pouvons noter est que les tailles de vocabulaire requises par ce nouveau type de vocabulaire se sont largement réduites. Les méthodes unigramme et SVM voient la taille de leur vocabulaire divisée par 2 et la TFIDF par 6.

4.2 Combinaison

Dans cette étude, nous avons présenté un ensemble de méthodes de détection de thème et nous avons pu remarquer que chacune de ces méthodes obtenait des performances maximales avec un vocabulaire qui lui était propre. De plus, celles-ci exploitent les données de façon différente. Afin d'améliorer les performances, nous envisageons d'exploiter les avantages de chacune de ces méthodes. Pour atteindre cet objectif, nous avons décidé de combiner ces différentes méthodes.

Pour les combiner, nous avons étudié un ensemble de méthodes : vote majoritaire, combinaison

linéaire, SVM et réseau de neurones. Nous présentons ici la méthode qui a permis d'obtenir les meilleurs résultats : le réseau de neurones (perceptron multi-couches).

Sur la couche d'entrée du perceptron, nous disposons de 35 valeurs (chaque méthode fournissant un score pour l'ensemble des 7 thèmes, et nous étudions 5 méthodes). La couche de sortie comporte 7 neurones, un pour chacun des thèmes possibles. Le perceptron utilisé comporte une seule couche cachée avec 15 neurones et exploite l'algorithme de rétropropagation. Afin de fixer les poids des neurones du perceptron, nous utilisons une méthode de validation croisée : nous avons divisé le corpus des 835 paragraphes en 7 sous-ensembles de tailles quasiment égales. Ensuite, nous avons effectué 7 tests : nous avons optimisé les paramètres du perceptron sur $6/7$ du corpus, puis nous avons évalué la qualité de ces valeurs sur $1/7$ (le reste) du corpus.

Après avoir effectué la combinaison, nous avons évalué le gain en performances obtenu pour les deux types de vocabulaires étudiés : tout d'abord les vocabulaires de l'état de l'art et ensuite les vocabulaires que nous proposons.

La combinaison des 5 méthodes en utilisant les vocabulaires définis dans l'état de l'art permet d'atteindre des performances en détection de thème de 87.2%, ce qui correspond à une amélioration des performances d'environ 5% (les performances les plus élevées dans ce cas étaient de 83.1%, modèle unigramme). Cette amélioration est statistiquement significative.

La méthode de contruction de vocabulaire que nous avons proposé ici a permis d'améliorer légèrement les performances (modèle unigramme : 83.4%). Lorsque nous combinons les méthodes avec ces vocabulaires, les performances atteignent 93.1%, ce qui correspond à une amélioration de près 11.6% des performances en détection de thème.

Nous pouvons conclure que la méthode de sélection de vocabulaire que nous proposons permet non seulement d'améliorer légèrement les performances en détection de thème. De plus, celle-ci amène à une réduction de la taille du vocabulaire nécessaire pour atteindre les résultats optimaux. Enfin, elle permet d'améliorer les performances en détection de thème de façon très conséquente (11.6%) lorsque les méthodes sont combinées.

Pour essayer de comprendre cette amélioration, nous nous sommes intéressés à la corrélation existant entre les scores fournis par les différentes méthodes, en fonction des vocabulaires utilisés. Nous avons pu remarquer que lorsque les vocabulaires utilisés étaient créés à l'aide de la méthode que nous proposons, la corrélation existant entre les méthodes est considérablement réduite, ce qui permet un gain potentiel plus élevé. Cependant, il serait intéressant d'étudier l'apport de chacune des méthodes dans le gain en performance, ce qui permettrait ensuite l'utilisation d'un sous ensemble de celles-ci.

5 Conclusion et perspectives

Dans cet article, nous avons présenté un ensemble de méthodes de détection de thème ainsi que des méthodes de sélection de vocabulaire. Nous avons évalué les performances obtenues par ces méthodes lorsqu'elles sont appliquées à nos données. L'ensemble de ces méthodes atteint des performances supérieures à 74%. La méthode la plus performante est le modèle unigramme avec 83.1%.

Après avoir montré l'importance de la taille du vocabulaire utilisé pour la détection de thème, nous avons présenté une nouvelle approche de la création de vocabulaires. Celle-ci passe par

l'exploitation de vocabulaires de thèmes. Ensuite, on procède à l'union de ces derniers pour obtenir le vocabulaire utilisé pour la détection de thème. L'utilisation de ces vocabulaires a tout d'abord montré deux avantages, elle permet non seulement d'améliorer les performances en détection de thèmes des méthodes étudiées, mais également la réduction de la taille du vocabulaire requise pour atteindre les performances maximales (facteur variant entre 2 et 6 en fonction des méthodes).

Dans l'optique d'améliorer les performances en détection de thème, nous avons ensuite étudié la combinaison des méthodes présentées. Nous avons cherché à combiner les résultats des méthodes dans le cas où les vocabulaires de la littérature sont exploités et également dans le cas où les vocabulaires utilisés sont ceux que nous proposons. Le gain obtenu dans le premier cas est important (5%). Celui constaté dans le cas de l'utilisation des vocabulaires que nous proposons est beaucoup plus grand puisqu'un gain de 11.6% a été obtenu.

Au vu du gain obtenu par la combinaison de méthodes, nous envisageons de nous intéresser à une autre façon de créer les vocabulaires utilisés pour la détection de thème. Jusqu'à présent, nous avons cherché les vocabulaires qui permettaient de maximiser les performances des méthodes indépendamment les unes des autres. Il serait peut-être intéressant de rechercher les vocabulaires qui permettent d'obtenir les performances maximales en combinant les méthodes, ces vocabulaires n'obtenant peut être pas les meilleures performances pour chacune des méthodes utilisées seules.

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.

FRAKES W. & BAEZA-YATES R. (1992). *Information Retrieval : Data Structures and Algorithms*. Prentice Hall, Englewood Cliffs,NJ.

JELINEK F. & MERCER R. (1980). Interpolated estimation of markov source parameters from sparse data. In *Proceedings of Workshop Pattern Recognition in Practice*, p. 381–397, Amsterdam.

LEWIS D. (1992). An Evaluation of Phrasal and Clustered Representation on a Text Categorization Task. In *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*, p. 37–50.

MCDONOUGH J., NG K., JEANRENAUD P., GISH H. & ROHLICEK J. (1994). Approaches to Topic Identification On The Switchboard Corpus. In *IEEE Transactions on Acoustics, Speech, and Signal Processing*, p. 385–388.

MITCHELL T. (1996). *Machine Learning*, chapter 3. Mc Graw Hill.

SALTON G. (1991). Developments in Automatic Text Retrieval. *Science*, **253**, 974–979.

SEYMORE K., CHEN S. & ROSENFELD R. (1998). Nonlinear Interpolation of Topic Models for Language Model Adaptation. In *Proceedings of the International Conference on Spoken Language Processing*, Sydney, Australia.

VAPNIK V. (1995). *The Nature of Statistical Learning Theory*. Spinger, New York.

YANG Y. & PEDERSEN J. (1997). A comparative study on feature selection in text categorization. In D. H. FISHER, Ed., *14th International Conference on Machine Learning, ICML-97*, p. 412–420, San Francisco, US : Morgan Kaufmann.