WSIM : une méthode de détection de thème fondée sur la similarité entre mots

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

Résumé - Abstract

L'adaptation des modèles de langage dans les systèmes de reconnaissance de la parole est un des enjeux importants de ces dernières années. Elle permet de poursuivre la reconnaissance en utilisant le modèle de langage adéquat : celui correspondant au thème identifié.
Dans cet article nous proposons une méthode originale de détection de thème fondée sur des vocabulaires caractéristiques de thèmes et sur la similarité entre mots et thèmes. Cette méthode dépasse la méthode classique (TFIDF) de 14%, ce qui représente un gain important en terme d'identification. Nous montrons également l'intérêt de choisir un vocabulaire adéquat. Notre méthode de détermination des vocabulaires atteint des performances 3 fois supérieures à celles obtenues avec des vocabulaires construits sur la fréquence des mots.

Speech recognition systems benefit from statistical language model adaptation, which is currently one of the most important challenge. This adaptation may go through the use of a particular language model : the one of the topic identified. In this article, a new and original topic identification method is presented, it is based on the similarity between words and topics. The performance of this method overcomes the one of reference, the TFIDF. The increase is 14% of topic identification.
The importance of the choice of topic vocabularies is also put forward. A judicious way to create them, instead of choosing the most probable words, makes performance triple.

Mots-clefs – Keywords

Reconnaissance de la parole, modélisation statistique du langage, détection de thème, information mutuelle, similarité.
Automatic speech recognition, statistical language modeling, topic detection, mutual information, similarity.

1 Introduction

Les modèles de langage (MLs) sont utilisés dans de nombreux domaines comme la reconnaissance de la parole, la traduction automatique, la recherche d'informations, la reconnaissance de l'écriture, etc. Les performances d'un système de reconnaissance automatique de la parole, notamment, sont fortement dépendantes des modèles de langage.
Un ML a pour but de modéliser le comportement de la langue. Ainsi, un ML statistique la représente sous la forme d'une distribution de probabilités de séquences de mots. Dans le cas d'un système de reconnaissance de la parole, le score fourni par le module acoustique, qui représente la correspondance entre le signal et une suite de mots donnée, est combiné avec celui fourni par le modèle de langage, qui représente la vraisemblance de cette même séquence. La séquence de mots correspondant au meilleur score sera celle retenue par le système.

Les modèles de langage les plus utilisés sont les modèles de type n-grammes, qui évaluent la probabilité d'un mot sachant les $n-1$ mots précédents. Le principal avantage de ces modèles réside dans leur simplicité de mise en œuvre. Cependant, lors de leur construction, on est souvent confronté à des problèmes de manque de données, résolus par l'utilisation de méthodes de lissage (*cf* (Chen & Goodman, 1996) pour une synthèse de ces méthodes). Plus la valeur de n est élevée, plus le problème du manque de données est important. Par conséquent, en pratique, la valeur de n excède rarement 3 (modèle trigrammes). Cependant, il est évident que la quantité d'information prise en compte par ces modèles pour évaluer la probabilité d'un mot est largement inférieure à celle qui joue effectivement un rôle lors de la prédiction d'un mot. Pour cette raison, de nombreux travaux ont été menés dans le but d'augmenter la taille de l'historique pris en compte : (Kuhn & De Mori, 1990) intègrent un cache au modèle de langage, ce qui a pour conséquence d'augmenter la probabilité des mots déjà apparus dans l'historique. Dans le même ordre d'idées, (Rosenfeld, 1996) y intègre des triggers de mots. Récemment, (Chelba & Jelinek, 2000) ont développé un modèle de langage qui combine un modèle n-grammes, un analyseur et un étiqueteur, permettant ainsi d'exploiter des mots apparaissant très loin dans l'historique.
Dans notre cas, nous travaillons dans l'hypothèse que le langage, ou plus exactement son vocabulaire caractéristique, varie en fonction du thème traité. Il est donc utile, toujours dans le but d'augmenter l'information prise en compte pour prédire un mot, de chercher à connaître le thème d'un texte pour ensuite adapter le ML à ce thème.
Nous nous intéressons tout particulièrement à cette phase de recherche du thème d'un texte. Dans cet article, nous proposons une nouvelle méthode de détection de thèmes, WSIM (Word SIMilarity), fondée non seulement sur la probabilité des mots dans les thèmes, unique information habituellement exploitée, mais également sur la similarité des mots avec les thèmes et l'utilisation de vocabulaires caractéristiques.

La section 2 explique en quoi les tâches de catégorisation de textes et détection de thèmes sont similaires puis présente un état de l'art des méthodes de catégorisation de textes. Nous introduisons, en section 3, les principes de notre méthode de détection de thèmes, dont les performances seront étudiées, et comparées à d'autres méthodes, en section 4. Enfin, nous conclurons et présenterons quelques perspectives.

2 La détection de thèmes

2.1 Définition

Soit un document d_i et $C = \{c_1, \ldots, c_J\}$ un ensemble de classes. La catégorisation de textes est la tâche qui consiste à assigner une ou plusieurs classes à d_i. Pour cela, nous disposons d'un corpus dit "d'apprentissage" composé d'un ensemble de documents dont la(es) classe(s) d'appartenance sont connues. Le système de détection de thème est ensuite entraîné sur cet ensemble d'apprentissage, dans le but de correctement catégoriser un nouveau document. Dans notre cas, une classe peut être assimilée à un thème, l'objectif étant de retrouver le thème c_k du document d_i.

2.2 Les travaux en catégorisation

Les méthodes classiques de catégorisation exploitent l'information contenue dans le document pour déterminer sa classe. Dans la majorité des cas, ce sont les mots qui sont utilisés pour représenter cette information.

Le problème de la catégorisation de textes a été largement étudié, nous en présentons ici plusieurs grandes approches, parmi les plus utilisées.
Le document est tout d'abord transformé sous la forme d'un vecteur où chaque élément représente "grossièrement" le poids d'un mot dans le document. Dans de rares cas, comme par exemple les arbres de décision binaires, ce vecteur contient des valeurs booléennes, représentant la présence ou non du mot dans le document : 1 si le mot est présent, 0 sinon, voir (Lewis & Ringuette, 1994).
Certaines méthodes sont fondées sur une approche probabiliste, elles évaluent la probabilité de chaque classe sachant le document donné. Le modèle unigramme thématique est l'exemple le plus connu de ces classifieurs (Mc Donough & Ng, 1994).
On peut à nouveau citer les arbres de décision (Mitchell, 1996). Chaque nœud représente un terme et chaque branche un test sur la fréquence de ce terme dans le document. Enfin les feuilles représentent une classe. La classe affectée au document est celle qui correspond à la feuille obtenue par parcours de l'arbre.
Dans l'approche par réseaux de neurones (Dagan *et al.*, 1997), le document à classer est présenté à l'entrée du réseau. La couche de sortie, quant à elle, représente l'ensemble des classes. Après activation du réseau, les valeurs de la couche de sortie représentent les classes possibles du document.
Enfin les Machines à Vecteur Support (SVMs), sont une famille de classifieurs qui minimise une borne supérieure sur l'erreur de généralisation. Elles sont fondées sur la séparation de données par hyperplan. Elles ont été appliquées à la catégorisation de textes dans (Joachims, 1998).
Les approches présentées ci-dessus ne sont pas exhaustives, on peut également citer les classifieurs à base de règles de décision (Apté *et al.*, 1994), à base de régression avec notamment le modèle LLSF (Yang & Chute, 1994), la méthode Rocchio (Joachims, 1997) avec tout particulièrement la TFIDF (Salton, 1991; Seymore & Rosenfeld, 1997), etc.

Certaines études ont également été menées en vue de l'exploitation d'informations de plus haut niveau que le mot. Ainsi, (Lewis, 1992) intègre des séquences de mots extraites en accord avec une grammaire, et (Caropreso *et al.*, 2001) des séquences de mots de nature purement

statistique. Les deux approches n'ont montré aucune amélioration des performances.

3 Description de WSIM

Dans cet article, nous proposons une nouvelle méthode de détection de thème, WSIM. Nous pouvons la classer dans la famille des méthodes probabilistes. Chaque thème est représenté par un vecteur, où chaque élément représente un mot. Contrairement aux méthodes classiques comme la TFIDF (Seymore & Rosenfeld, 1997), les éléments du vecteur ne représentent pas uniquement le poids du mot dans le thème, celui-ci est combiné avec leur "similarité". La similarité entre un mot x et un thème T_j est fondée sur la similarité entre ce mot et l'ensemble des mots caractéristiques du thème T_j.

3.1 La mesure de similarité entre deux mots

(Dagan *et al.*, 1999) introduit une mesure de similarité entre 2 mots x et y, évaluée en se basant sur leurs comportements respectifs en contexte (droit et gauche). Plus précisément, deux mots sont considérés comme similaires si leurs informations mutuelles avec l'ensemble des autres mots du vocabulaire sont proches. Cette similarité est évaluée de la manière suivante :

$$Similarite(x,y) = \frac{1}{2V}\sum_{i=1}^{|V|}\frac{min\left(I(z_i,x),I(z_i,y)\right)}{max\left(I(z_i,x),I(z_i,y)\right)} + \frac{min\left(I(x,z_i),I(y,z_i)\right)}{max\left(I(x,z_i),I(y,z_i)\right)} \qquad (1)$$

Où V est le vocabulaire et $I(z_i,x)$ est l'information mutuelle entre les mots z_i et x.
Cette mesure a été initialement développée dans le but d'estimer la probabilité de cooccurrences de mots, non observées à l'apprentissage. Nous avons adopté cette mesure pour développer une méthode permettant d'identifier le thème d'un document. Cette méthode est fondée sur l'information mutuelle I calculée sur une fenêtre glissante de d mots, la nature de la similarité est donc plus sémantique que syntaxique. $I(z_i,x)$ est évaluée de la manière suivante :

$$I(z_i,x) = P_d(z_i,x)\log\frac{P_d(z_i,x)}{d^2.P(z_i)P(x)}$$

où d représente la distance ou la taille de la fenêtre glissante. $P_d(z_i,x)$ est la probabilité de succession des mots z_i et x à une distance au plus d. $P(x)$ représente la probabilité *a priori* du mot x.
Toujours dans un but de détection de thème, nous ne cherchons pas à connaître la similarité entre deux mots dans le langage, mais dans un thème donné. Par conséquent, la similarité entre deux mots, pour le thème T_j, sera évaluée comme suit :

$$Similarite_j(x,y) = \frac{1}{2l_j}\sum_{i=1}^{|l_j|}\frac{min\left(I_j(z_i,x),I_j(z_i,y)\right)}{max\left(I_j(z_i,x),I_j(z_i,y)\right)} + \frac{min\left(I_j(x,z_i),I_j(y,z_i)\right)}{max\left(I_j(x,z_i),I_j(y,z_i)\right)} \qquad (2)$$

où l_j est le vocabulaire du thème T_j, et $I_j(z_i,x)$ est évaluée sur le corpus d'apprentissage du thème T_j.

TAB. 1 – Label des thèmes étudiés et leur taille d'apprentissage

Thème	Nombre de mots d'apprentissage	Thème	Nombre de mots d'apprentissage
Culture	25 M	Politique	13 M
Économie	21 M	Sciences	2 M
Étranger	24 M	Sports	170 K
Histoire	560 K		

3.2 La similarité entre un mot et un thème

Soit $V_j = v_{jx_1}, v_{jx_2}, \ldots, v_{jx_{|l_j|}}$ le vecteur représentant le thème T_j. Chaque élément du vecteur représente la similarité entre un mot et le thème.
Nous proposons d'estimer la similarité entre le mot x et le thème T_j comme étant la moyenne des similarités entre le mot x et les mots du vocabulaire caractéristique de T_j. Cette dernière est ensuite pondérée par la probabilité du mot dans le thème :

$$v_{jx} = Sim(x, T_j) = P(x \mid T_j) \frac{\sum_{k=1}^{|l_j|} Similarite_j(x, y_k)}{\sum_{x=1}^{|l_j|} \sum_{k=1}^{|l_j|} Similarite_j(x, y_k)} \quad (3)$$

3.3 La détermination du thème d'un document

En phase de test, pour chaque thème T_j ($j \in 1..J$), nous disposons d'un vecteur V_j. Le score de chaque thème sachant le document de test d composé de N mots $d = w_1, w_2, \ldots, w_N$ est évalué comme étant la somme normalisée des similarités entre les mots du document et le thème :

$$P(T_j \mid d) = \varphi_j \frac{\sum_{i=1}^{N} v_{jw_i}}{\sum_{k=1}^{J} \sum_{i=1}^{N} v_{kw_i}} * \sum_{i=1}^{N} \delta_{ij} \quad (4)$$

avec $\delta_{ij} = \begin{cases} 1 & \text{si } w_i \in l_j \\ 0 & \text{sinon} \end{cases}$ et $\sum_{i=1}^{N} \delta_{ij}$ représente le nombre de mots de l_j dans d, φ_j est un coefficient de pondération thématique avec $\sum_{j=1}^{J} \varphi_j = 1$. Les valeurs de φ_j sont déterminées par validation croisée, sur un corpus d'optimisation. Finalement, le thème retenu est celui qui maximise (4).

4 Expérimentations

4.1 Les données

Les expériences de détection de thèmes sont évaluées sur un corpus issu du journal *Le Monde*, des années 1987 à 1991 (plus de 80 M mots). Ce corpus est divisé en 7 thèmes, inégalement représentés. La liste des thèmes ainsi que leur taille d'apprentissage sont présentées TAB. 1. Ce corpus est disponible sous forme d'articles. Cependant, à l'intérieur d'un même article, on peut être confronté à des changements de thèmes, problème auquel nous ne nous intéressons pas. Par

conséquent, nous avons extrait aléatoirement 835 paragraphes, que nous considérons ne traiter que d'un seul thème, et qui forment le corpus de test.

4.2 Construction du vocabulaire

Dans l'équation (2), la similarité entre deux mots x et y pour le thème T_j se calcule sur le vocabulaire du thème, qu'il nous faut donc construire. Comme le montrent de nombreuses études (Brun *et al.*, 2000; Mladenic, 1998), le vocabulaire d'un thème constitue le noyau de base sur lequel repose toute méthode d'identification. Par conséquent, il est indispensable de ne pas se contenter des mots les plus fréquents des thèmes, mais d'en trouver les termes caractéristiques. Nous étudions ici deux méthodes de construction des vocabulaires de thèmes : une méthode classique et une méthode adaptée à la catégorisation de textes/détection de thèmes.

4.2.1 Mots les plus fréquents de chaque thème

Le premier ensemble de vocabulaires étudié est construit de façon très classique, où chaque vocabulaire de thème contient les n mots les plus fréquents (en absolu) du corpus d'apprentissage de ce thème. Les mots outils (non porteurs de sens) ont évidemment été supprimés.

4.2.2 Information mutuelle mot-thème

Le second ensemble de vocabulaires de thèmes est construit d'une manière plus judicieuse. Il s'agit d'évaluer la quantité d'information apportée par la variable T (thème) à la variable X (mot). Cette quantité est mesurée à l'aide de l'information mutuelle :

$$I(x, T_j) = P(x, T_j) \log \frac{P(x, T_j)}{P(x)P(T_j)} \tag{5}$$

avec $P(x, T_j)$ est la probabilité conjointe d'apparition de x et T_j, $P(x)$ est la probabilité *a priori* du mot x et $P(T_j)$ la probabilité *a priori* du thème T_j.
Une information mutuelle élevée entre un mot et un thème est le signe d'un lien fort entre ces deux éléments. Par conséquent, les vocabulaires de thèmes seront composés des mots d'information mutuelle les plus élevées.

Le nombre de mots par thème doit maintenant être fixé. Nous choisissons volontairement un nombre de mots identique pour chaque thème. Concernant la méthode exploitant l'information mutuelle, afin de déterminer le nombre de mots à conserver, nous trions, pour chaque thème, les mots par ordre décroissant de valeur d'information mutuelle avec le thème, et nous traçons la courbe de cette évolution (FIG. 1). A l'aide de cette courbe, nous pouvons remarquer que globalement, au dessus de 2000 mots, l'information mutuelle des mots se "stabilise", nous choisissons donc de conserver 2000 mots par thème.

Dans un souci de rigueur de comparaison des méthodes, nous avons également fixé n à 2000 pour les vocabulaires composés des mots les plus fréquents.

Nous avons évalué les performances de WSIM sur chacun des deux vocabulaires. Les résultats sont présentés TAB. 2. La différence spectaculaire des résultats peut être expliquée par des vocabulaires de thèmes très différents : en effet, en moyenne les deux ensembles de vocabulaires

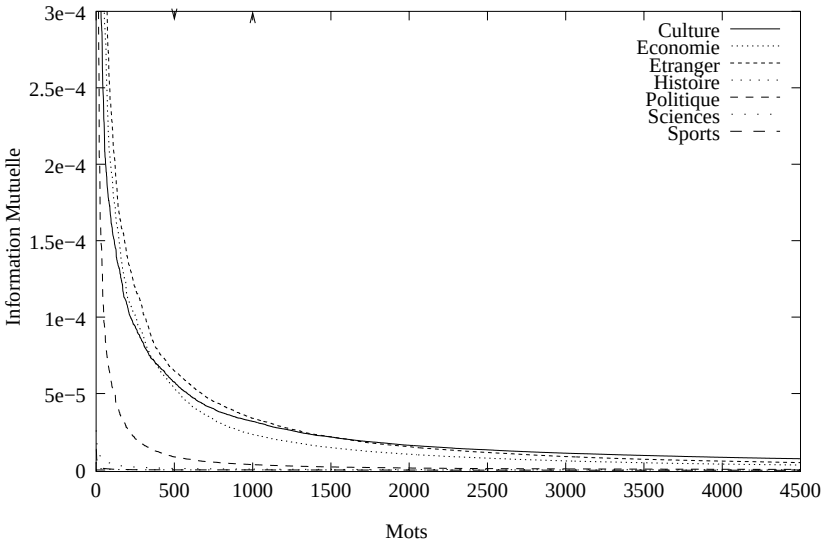

FIG. 1 – Information mutuelle classée par valeur décroissante pour chaque thème

TAB. 2 – Taux de détection des thèmes en fonction du vocabulaire

Vocabulaire	Performance (%)
Plus fréquents	27.5
Information mutuelle	82.4

ont seulement 31% des mots en commun. De plus, les vocabulaires construits avec la méthode des n plus fréquents, ont un taux de recouvrement de 64%, alors que pour les vocabulaires construits avec l'information mutuelle, ce taux n'est que de 25%.
Nous pouvons ainsi noter que la méthode WSIM semble être très dépendante du vocabulaire choisi. En effet, la similarité entre deux mots est fondée sur leur comportement avec l'ensemble des autres mots du vocabulaire. Par conséquent, les mots du thème doivent être particulièrement bien choisis pour obtenir une similarité fiable.
Nous décidons donc de conserver les vocabulaires créés à l'aide de la mesure d'information mutuelle.

4.3 Résultats

Afin d'étudier les performances de WSIM, nous proposons de la comparer avec d'autres méthodes de détection de thèmes.
Nous choisissons pour cela, de la comparer à la TFIDF (Salton, 1991), qui est la méthode citée comme référence dans le domaine. Cette dernière évalue la distance cosine entre les distributions de probabilités des mots dans les thèmes et celle du document de test.
Nous comparons également les performances (en termes de rappel) de notre méthode à celle du modèle cache (Bigi *et al.*, 2000). Les expériences menées dans une étude récente (Bigi *et al.*, 2001), et qui comparait 5 méthodes donnait la méthode d'identification par cache en tête. La méthode cache évalue la distance de Kullback-Leibler entre les distributions de mots des thèmes et celle du cache du document de test.

TAB. 3 – Performance des trois méthodes étudiées

Méthode	Performance (%)
TFIDF	72.1
Cache	82.0
WSIM	82.4

TAB. 4 – Rappel, précision et F_1 pour chaque méthode et chaque thème

	TFIDF			Cache			WSIM		
Thème	Rap	Prec	F_1	Rap	Prec	F_1	Rap	Prec	F_1
Culture	83.2	82.3	82.8	84.7	**90.4**	**87.4**	**85.3**	87.1	86.2
Economie	60.8	89.8	72.4	74.6	**91.0**	**82**	**78.8**	84.6	81.6
Etranger	57.8	58.4	58.2	**86.3**	73.9	79.6	85.3	**79.8**	**82.5**
Histoire	**58.3**	13.2	21.6	16.6	14.3	**14.4**	8.3	**33.3**	13.3
Politique	70.7	79.5	74.8	85.1	75.1	79.8	**86.2**	**83.0**	**84.6**
Sciences	**90.8**	66.7	76.9	88.1	**82.7**	**85.4**	83.5	79.8	81.6
Sports	66.7	59.3	62.8	75	**72**	**73.4**	**83.3**	62.5	71.4

Dans TAB. 3, nous comparons les performances de notre méthode à la TFIDF et au cache sur l'ensemble des 835 paragraphes. Les performances de la TFIDF, qui est la méthode de référence, sont largement dépassées par les deux autres méthodes. De plus, les performances du modèle cache ont été dépassées, pour la première fois, par la méthode WSIM, même si leurs performances restent cependant très proches.
Afin de mieux analyser les performances de ces 3 méthodes, il serait donc intéressant de les étudier thème par thème. Pour cette raison, nous présentons les performances de chacune des méthodes en termes de rappel et précision, où :

$$\text{Rappel}_T = \frac{\text{Nb textes correctement étiquetés T}}{\text{Nb textes d'étiquette T}} \tag{6}$$

$$\text{Précision}_T = \frac{\text{Nb textes correctement étiquetés T}}{\text{Nb textes étiquetés T}} \tag{7}$$

Rappelons ici que l'objectif final de la détection de thèmes est l'adaptation du modèle de langage au thème. Par conséquent, nous cherchons une méthode qui détecte à la fois le thème du plus grand nombre de documents (rappel) mais fournisse également une étiquette fiable (précision).
Pour cette raison, les résultats seront également présentés en termes de mesure F_1, qui permet de combiner le rappel et la précision dans une seule valeur.

$$\text{F}_1 = \frac{2 * \text{rappel}_T * \text{précision}_T}{\text{rappel}_T + \text{précision}_T} \tag{8}$$

TAB. 4 présente, par thème, les valeurs de rappel, précision et F_1 pour les 3 méthodes étudiées.

Nous pouvons remarquer que les performances des méthodes varient significativement d'un thème à l'autre. Une des raisons de ces différences peut être la taille d'apprentissage des thèmes, les thèmes bien appris ayant tendance à être bien détectés. En effet, comme présenté dans la

table 1, la taille d'apprentissage entre *Culture* et *Sports* diffère d'un facteur environ 150. Une autre raison peut également être le recouvrement entre thèmes.
Les performances par thème de la TFIDF sont représentatives de son comportement général : seules ses valeurs de rappel pour les thèmes *Histoire* et *Sciences* surpassent les deux autres méthodes.

On peut remarquer que le thème *Histoire* n'est bien reconnu par aucune des méthodes. En plus d'une taille d'apprentissage faible, le thème *Histoire* souffre d'un manque de vocabulaire propre. En effet, on peut intuitivement dire que le thème *Histoire* ne peut être représenté à l'aide de mots de vocabulaires spécifiques. Ce dernier est plutôt représenté par des dates (ensemble quasi-infini) ou encore par l'emploi d'un temps passé. Il serait donc intéressant d'intégrer des informations d'un niveau supérieur au mot et éventuellement des connaissances syntaxiques pour améliorer les performances de détection.
Bien que la méthode WSIM obtienne les meilleures performances sur le corpus général, on peut remarquer que le modèle cache la dépasse légèrement dans certains cas, et notamment au niveau de la précision. Cependant, WSIM a un taux de rappel beaucoup plus homogène sur l'ensemble des thèmes, hors *Histoire,* que le modèle cache, ses performances ne semblant pas dépendre de la taille d'apprentissage.

5 Conclusions et perspectives

Nous avons présenté une nouvelle méthode de détection de thème WSIM, originale par l'information qu'elle exploite. En plus de la probabilité des mots dans les thèmes, celle-ci utilise la similarité mot-thème, elle-même basée sur la similarité inter-mots.
Cette méthode est très dépendante du vocabulaire utilisé en raison de l'utilisation de l'intégralité des mots des vocabulaires pour calculer la similarité inter-mots. Nous avons montré l'importance du choix des vocabulaires : nos tests donnent des performances 3 fois supérieures à celles obtenues avec un vocabulaire composé des mots les plus fréquents.
Comparée à la méthode TFIDF classique, la méthode WSIM obtient des résultats meilleurs de 14%. Les performances en détection de thèmes dépassent également celles du modèle cache, qui n'avait jusque là jamais été égalé. Même si cette amélioration n'est pour le moment pas spectaculaire, nous travaillons actuellement sur plusieurs pistes en vue d'améliorer nos résultats.
Nous avons volontairement choisi des vocabulaires de thèmes composés d'un nombre égal de mots. Cependant, les vocabulaires des thèmes avec une taille d'apprentissage faible, comportent vraisemblablement des mots non caractéristiques. Il serait donc intéressant d'étudier des vocabulaires avec des tailles différentes, en fixant par exemple une valeur d'information mutuelle normalisée minimale.
De plus, malgré des études prouvant la non amélioration des performances de détection en utilisant des séquences de mots, nous envisageons d'insérer des séquences de mots. Tout d'abord en raison de la taille d'apprentissage de nos thèmes, qui est largement supérieure à celle des études citées précédemment. De plus, la façon dont nous souhaitons procéder pour l'extraction de séquences est fondée sur des connaissances sémantiques.
Pour le traitement de thèmes comme *Histoire*, il serait intéressant d'utiliser des classes sémantiques, pour notamment représenter des notions de dates, lieux, noms propres, etc., le vocabulaire devenant ainsi un vocabulaire de classes.

Références

APTÉ G., DAMERAU F. & WEISS S. (1994). Automated learning of decision rules for text categorization. *ACM Transactions on Information Systems*, **12**(3), 233–251.

BIGI B., BRUN A., HATON J., SMAILI K. & ZITOUNI I. (2001). Dynamic topic identification : Towards combination of methods. In *Proceedings of the Recent Advances in Natural Language Processing (RANLP)*, p. 255–257.

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

BRUN A., SMAILI K. & HATON J. (2000). Experiment analysis in newspaper topic detection. In *7th International Symposium on String Processing and Information Retrieval, SPIRE-2000*, p. 55–64.

CAROPRESO M., MATWIN S. & SEBASTIANI F. (2001). *A learner-independant evaluation of the usefulness of statistical phrases for automatic text categorization*, p. 78–102. Hershey, US.

CHELBA C. & JELINEK F. (2000). Structured language modeling. *Computer Speech and Language*, **14**(4), 283–332.

CHEN S. F. & GOODMAN J. (1996). An empirical study of smoothing techniques for language modeling. In *Proceedings id the 34th Annual Meeting of the ACL*, p. 310–318.

DAGAN I., KAROV Y. & ROTH D. (1997). Mistake-driven learning in text categorization. In *Proceedings of the EMNLP-97, 2nd Conference on Empirical Methods in Natural Language Processing*, p. 55–63.

DAGAN I., LEE L. & PEREIRA F. C. N. (1999). Similarity-based models of word coocurrence probabilities. *Machine Learning*, **34**, 43–69.

JOACHIMS T. (1997). A probabilistic analysis of the rocchio algorithm with tfidf for text categorization. In *Proceedings of ICML-97, 14th International Conference on Machine Learning*, p. 143–151.

JOACHIMS T. (1998). Text categorization with support vector machines : learning with many relevant features. In *Proceeding of ECML-99, 16th European Conference on Machine Learning*, p. 137–142.

KUHN R. & DE MORI R. (1990). A cache-based natural language model for speech reproduction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **12**(6), 570–583.

LEWIS D. (1992). An evaluation of phrasal and clustered representations on a text categorization task. In A. PRESS, Ed., *Proceedings of SIGIR-92, 15th ACM International Conference on Research and Development in Information Retrieval*, p. 37–50, New York, US.

LEWIS D. & RINGUETTE M. (1994). A comparison of two learning algorithms for text categorization. In *Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval*, p. 81–93.

MC DONOUGH J. & NG K. (1994). Approaches to topic identification on the switchboard corpus. In *International Conference on Acoustics, Speech and Signal Processing*, p. 385–388, Yokohama, Japan.

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

MLADENIC D. (1998). Feature subset selection in text-learning. In *10th European Conference on Machine Learning ECML98*.

ROSENFELD R. (1996). A maximum entropy approach to adaptive statistical language modeling. *Computer Speech and Language*, **10**, 187–228.

SALTON G. (1991). Developments in automatic text retrieval. *Science*, **253**, 974–979.

SEYMORE K. & ROSENFELD R. (1997). *Large-scale Topic Detection And Language Model Adaptation*. Rapport interne CMU-CS-97-152, School of Computer Science,CMU.

YANG Y. & CHUTE C. (1994). An example-based mapping method for text categorization. *ACM Transactions on Information Systems*, **12**(3), 252–277.