Un modèle d'espace vectoriel de concepts pour noyaux sémantiques

Sujeevan Aseervatham*

*LIPN - UMR 7030
CNRS - Université Paris 13
99, Av. J.B. Clément
F-93430 Villetaneuse, France
Sujeevan.Aseervatham@lipn.univ-paris13.fr

Résumé. Les noyaux ont été largement utilisés pour le traitement de données textuelles comme mesure de similarité pour des algorithmes tels que les Séparateurs à Vaste Marge (SVM). Le modèle de l'espace vectoriel (VSM) a été amplement utilisé pour la représentation spatiale des documents. Cependant, le VSM est une représentation purement statistique. Dans ce papier, nous présentons un modèle d'espace vectoriel de concepts (CVSM) qui se base sur des connaissances linguistiques a priori pour capturer le sens des documents. Nous proposons aussi un noyau linéaire et un noyau latent pour cet espace. Le noyau linéaire exploite les concepts linguistiques pour l'extraction du sens alors que le noyau latent combine les concepts statistiques et linguistiques. En effet, le noyau latent utilise des concepts latents extraits par l'Analyse Sémantique Latente (LSA) dans le CVSM. Les noyaux sont évalués sur une tâche de catégorisation de texte dans le domaine biomédical. Le corpus Ohsumed, bien connu pour sa difficulté de catégorisation, a été utilisé. Les résultats ont montré que les performances de catégorisation sont améliorées dans le CSVM.

1 Introduction

Les mesures de similarité sont des éléments clés dans les algorithmes de traitement automatique des langues. Elles sont utilisées pour orienter le processus d'extraction de connaissance. Ainsi, elles sont les principales responsables des performances d'un algorithme. Si une mesure de similarité pertinente améliorera les performances, une mauvaise mesure risque de mener à des résultats incohérents. La définition d'une bonne mesure n'est pas un processus aisé. En effet, la mesure doit donner une bonne indication sur le degré de similarité entre deux documents. La notion de sémantique n'est pas clairement définie. Bien que nous essayons d'imiter la perception humaine, l'information sémantique peut prendre différente forme selon l'approche adoptée. Il existe deux grandes approches : l'une basée sur l'information statistique tel que la fréquence de co-occurrence des termes et l'autre basée sur des sources de connaissances externes telles que les ontologies.

Dans la communauté de l'apprentissage, les noyaux (Shawe-Taylor et Cristianini, 2004) sont utilisés depuis une décennie comme fonctions de similarité basées sur le cosinus formé

par deux vecteurs. Les noyaux sont, en réalité, des produits scalaires définis dans un espace de Hilbert. Ils ont la spécificité de plonger, implicitement, les données dans un espace de Hilbert, dit espace de description, avant de calculer le produit scalaire. Les noyaux peuvent être intégrés dans tout algorithme d'apprentissage basé sur le produit scalaire tel que les Séparateurs à Vaste Marge (SVM) (Vapnik, 1995). Ainsi, ils étendent l'utilisation de l'apprentissage numérique aux tâches du TAL (Traitement Automatique des Langues). En effet, aucune contrainte n'étant imposée sur l'espace des données, les noyaux peuvent être définis pour tout type de données tel que les données textuelles. Le modèle d'espace vectoriel (VSM) (Salton et al., 1975), représentant un document sous la forme d'un vecteur de fréquences de termes, est largement utilisé. Les noyaux basés sur ce modèle ont permis d'obtenir des résultats très prometteurs dans le domaine de la catégorisation de texte (Joachims, 2002, 1998).

Dans cet article, nous présentons un modèle d'espace vectoriel de concepts (CVSM) pour la représentation des documents textuels. Cette représentation, induite par des connaissances a priori, est présentée, ici, comme une alternative au modèle classique d'espace vectoriel. Le VSM se base uniquement sur la fréquence d'occurrence des termes. Pour le CVSM, une taxonomie de concepts linguistiques est utilisée comme source de connaissance pour définir l'espace vectoriel. De plus, nous proposons deux noyaux basés sur les concepts. Le premier noyau, le noyau CVSM linéaire, est défini dans le CVSM où chaque document est représenté par un vecteur de concept intégrant l'information sur la taxonomie des concepts. Le second noyau, le noyau CVSM latent, mélange l'approche agnostique basée sur l'information statistique et l'approche a priori utilisant des connaissances externes propres au domaine. Basé sur le noyau LSA (Cristianini et al., 2002), le noyau CVSM latent utilise une décomposition en valeurs singulières (SVD) pour découvrir des structures latentes entre les concepts linguistiques du CVSM.

L'utilisation des noyaux CVSM est illustrée par une tâche de catégorisation de texte dans le domaine biomédical. L'*Unified Medical Language System* (UMLS) est utilisé tant que source a priori de connaissances biomédicales pour l'extraction de concepts à partir documents textuels. Les performances de ces noyaux sont évaluées sur cette tâche en utilisant un classifieur SVM. Le corpus Ohsumed qui est connu pour être un corpus difficile, est utilisé pour l'évaluation expérimentale.

2 Le modèle d'espace vectoriel (VSM)

Le modèle de l'espace vectoriel (VSM) est la représentation des documents textes dans un espace vectoriel la plus communément utilisée. Il a été introduit dans (Salton et al., 1975) pour le problème d'indexation des documents. Dans le VSM, un document est associé à un vecteur où chacun de ses composants représente la fréquence d'occurrence d'un terme dans le document. Le VSM est, ainsi, doté d'un repère où chaque axe représente la fréquence d'un terme précis. Du fait de sa simplicité et de son efficacité, le VSM a été très largement adopté pour résoudre une large variété de problème en TAL. Dans (Joachims, 2002, 1998), plusieurs classifieurs tels les SVM et les K-NN ont été utilisés dans le VSM pour des problèmes de catégorisation de texte. Ces classifieurs, et en particulier les SVM, ont donné d'excellents résultats dans le VSM.

Pour raffiner le VSM, l'Analyse Sémantique Latente (LSA) (Deerwester et al., 1990) a été proposée. L'idée principale est de représenter un document dans le VSM non pas par ses termes mais par ces concepts. L'hypothèse est que les concepts sont plus adaptés pour modéliser le sens des documents que les termes. Dans la LSA, un espace sémantique de faible dimension est défini en appliquant une Décomposition en Valeurs Singulières (SVD) sur la matrice de termes par document. Les vecteurs documents du VSM sont alors projetés dans l'espace sémantique. En utilisant ce nouvel espace, un Noyau Sémantique Latent a été proposé dans (Cristianini et al., 2002). Ce noyau a été utilisé pour projeter les documents du VSM vers l'espace sémantique et calculer le produit scalaire. Le noyau a été utilisé avec succès sur une tâche de catégorisation de texte. Il a été montré que le noyau LSA peut atteindre les mêmes performances, dans un espace de très faible dimension, que le noyau du VSM utilisé dans (Joachims, 2002, 1998)).

Bien que l'espace sémantique de la LSA permet d'obtenir de très bonnes performances, l'espace étant défini par des concepts statistiques, il est très difficile, linguistiquement, d'interpréter cet espace et les concepts.

3 Noyau linéaire du modèle d'espace vectoriel de concepts

3.1 Pré-traitement des documents

Les documents textuels sont formatés pour une utilisation humaine. Ainsi, ils contiennent une riche variété de symboles et de protocoles tels que les règles typographiques. Ces informations sont ajoutées pour rendre la lecture et la compréhension plus aisées. Toutefois, le traitement automatique de ces données est rendu compliqué. En effet, si les éléments d'un document ne sont pas correctement gérés, ils peuvent être une source de bruits et d'ambiguïtés qui entraînera inexorablement une baisse des performances du système. Une façon d'éviter ce problème est de pré-traiter le document pour avoir une représentation adaptée au système de traitement.

Dans le cadre de ce travail, nous utiliserons le pré-traitement suivant :

1. Nettoyage du document : Tous les éléments qui peuvent introduire du bruit sont éliminés. Ainsi, tous les nombres, aussi bien sous un format numérique que sous un format littéral et les données non-textuelles sont retirés. Il est à noter que nous utilisons le terme "bruit" dans un sens général.

2. Segmentation du texte : Le document est segmenté en unité lexicale composée d'un ou plusieurs mots. Pour cela, un lexique doit être utilisé. Dans notre, cas, le lexique médical "*The Specialist Lexicon*" de l'UMLS a été utilisé.

3. Suppression des mots vides : Les unités lexicales ne contenant que des mots vides, à savoir, sans signification sont éliminés.

4. Normalisation des termes : Pour éviter les différentes formes fléchies des mots, un processus de normalisation est effectué à l'aide d'un lexique. La normalisation consiste à lemmatiser chaque mot d'une unité lexicale et à les ordonner alphabétiquement au sein de cette unité. Les unités lexicales absentes du lexique sont décomposées en unité lexicale composée d'un seul mot. Un processus de *stemming* est ensuite appliqué à chacun de ces mots.

5. Annotation Sémantique : Chaque unité lexicale est associée à un groupe de concepts (composé d'un ou plusieurs concepts). Cette étape nécessite l'utilisation d'un thésaurus. En outre,

une taxonomie de relation "est-un" entre les concepts sera utilisée lors de la phase de traitement. Dans le cadre de notre travail, le *Metathesaurus* de l'UMLS a été utilisé. Ce thésaurus intègre une ontologie de concepts complexe. L'ontologie a été transformée en taxonomie en ne conservant que les relations "est-un" et en supprimant les cycles.

3.2 Le modèle d'espace vectoriel de concepts

Le modèle d'espace vectoriel (VSM) est basé sur la forme morphologique des termes. Il est, ainsi, hautement dépendant de la langue du texte et de la fréquence d'occurrence des termes. En effet, le VSM utilise simplement la fréquence des termes pour capturer l'information d'un document. Il est, ainsi, limité par le fait qu'il ne peut correctement gérer les termes synonymes et les termes polysémiques. De plus, les liens entre les termes sémantiquement proches ne peuvent être modélisés. Toutefois, l'espace à haute dimension induit par le VSM permet aux systèmes, utilisant le VSM, d'obtenir des performances qui sont parmi les meilleures (Joachims, 1998).
Dans cette section, nous présentons le modèle d'espace vectoriel de concepts (CVSM), basé sur le VSM. Ce modèle devrait permettre de gérer les problèmes, rencontrés par le VSM, énumérés ci-dessus. Le CVSM est un espace vectoriel dans lequel chaque axe représente un concept défini dans un dictionnaire de concepts. Le dictionnaire est constitué des concepts définis dans un thésaurus et des mots racines, obtenus par *stemming*, lorsque ces mots ne peuvent être associés à des concepts. Nous prenons, alors, l'hypothèse que ces mots expriment un concept à part entière. Dans le CVSM, les documents sont pré-traités selon la méthode décrite plus haut. Une fois qu'un document d a été pré-traité, chaque unité lexicale l du document d est associée à un vecteur de concepts local $\phi(l)$. Le $i^{\text{ème}}$ composant $\phi_i(l)$ de $\phi(l)$ associé au concept c_i est alors donné par :

$$\phi_i(l) = \sum_{t \in \mathcal{N}(l)} \frac{\phi_i^n(t)}{||\phi^n(t)||} \tag{1}$$

où $\mathcal{N}(t)$ est l'ensemble de termes normalisés d'une unité lexicale l et $\phi_i^n(t)$ et le $i^{\text{ème}}$ composant, associé au concept c_i, du vecteur de concept $\phi^n(t)$ pour le terme normalisé t. $\phi_i^n(t)$ est défini par :

$$\phi_i^n(t) = \sum_{c \in \mathcal{C}(t)} \frac{\sum_{p \in \mathcal{P}(c)} \sigma_p(c, c_i)}{\sqrt{\sum_j (\sum_{p \in \mathcal{P}(c)} \sigma_p(c, c_j))^2}} \tag{2}$$

où

$$\sigma_p(c_i, c_j) = \begin{cases} (d_p(c_i, c_j) + 1)^{-\alpha} & \text{si } c_i, c_j \in p\,, \\ 0 & \text{sinon}\,, \end{cases} \tag{3}$$

$\mathcal{C}(t)$ est l'ensemble des concepts associés à t, $\mathcal{P}(c)$ est l'ensemble des chemins allant du concept c au concept racine dans la taxonomie, c'est à dire les chemins allant du concept spécifique c au concept le plus général, $d_p(c_i, c_j)$ est la distance entre le concept c_i et le concept c_j sur le chemin p et $\alpha \in [0, 1]$ est une valeur exprimant la puissance de décroissance. α est utilisé pour décroître l'influence des concepts généraux sur la représentation d'un terme. En effet, étant donné un terme t, un concept c associé à t, et p un chemin allant de c au concept le plus

général (le concept racine), les concepts proches, en terme de distance, de c fourniront le sens principal de t comparés aux concepts éloignés de c. Les concepts généraux n'expriment pas clairement le sens pertinent de t et peuvent même mener à des ambiguïtés voire du bruit. Par conséquent, il peut être intéressant de diminuer le pouvoir expressif des concepts selon leurs distances par rapport au concept spécifique c. En fixant α à un, nous pouvons considérablement réduire l'expressivité des concepts généraux et en fixant α à zéro, nous pouvons donner le même pouvoir d'expression à tous les concepts en ne faisant aucune différence entre les concepts généraux et spécifiques. α doit être fixé de manière empirique en fonction du corpus.

Étant donné un ensemble de vecteurs de concepts locaux $\{\phi(l_1), \ldots, \phi(l_n)\}$ pour un document d, le vecteur de concept global $\mathbf{\Phi}(d)$ pour d est donné par la formule suivante :

$$\mathbf{\Phi}(d) = \sum_{i=1}^{n} \frac{1}{||\phi(l_i)||}.\phi(l_i) \tag{4}$$

La normalisation dans l'équation 4 est utilisée pour représenter des documents de longueur différente avec une même échelle. De plus, les concepts qui apparaissent dans beaucoup de documents du corpus, ne fourniront pas d'information utile pour la discrimination de documents alors que les concepts rares dans un corpus peuvent être significatifs. Par conséquent, nous proposons d'utiliser un vecteur pondéré $\mathbf{\Phi}^{IDF}(d)$ en utilisant la pondération IDF (*Inverse Document Frequency*). Ce vecteur est défini tel que :

$$\Phi_i^{IDF}(d) = -\log(\frac{card(\mathcal{D}(c_i))}{N}).\Phi_i(d) \tag{5}$$

où N est le nombre de document dans le corpus et $\mathcal{D}(c_i)$ est l'ensemble des documents du corpus contenant le concept c_i.

La figure 1 illustre la modélisation d'un document textuel dans le CVSM.

3.3 Le noyau linéaire

Nous définissons le noyau linéaire dans le modèle d'espace vectoriel de concept comme ceci :

$$k_{CVSM}(d_1, d_2) = \frac{\langle \mathbf{\Phi}^{IDF}(d_1), \mathbf{\Phi}^{IDF}(d_2) \rangle}{||\mathbf{\Phi}^{IDF}(d_1)||.||\mathbf{\Phi}^{IDF}(d_2)||} \tag{6}$$

4 Le noyau CVSM latent

L'analyse sémantique latente permet de mettre à jour des relations entre des termes. Ces relations sont extraites par une décomposition linéaire en valeurs singulières de la matrice des fréquences de termes par document. Elles sont par conséquent des relations de co-occurrences des termes dans les documents. Les relations extraites sont appelés "concepts latents". La LSA permet, ainsi, non seulement de diminuer la dimension de l'espace en ne conservant que les relations les plus importantes mais aussi de gérer la polysémie et la synonymie.

Dans un même esprit, nous proposons, dans ce papier, l'utilisation de la LSA dans le CVSM pour capturer les structures latentes entre les concepts. Les concepts LSA peuvent être perçus

Un modèle d'espace vectoriel de concepts

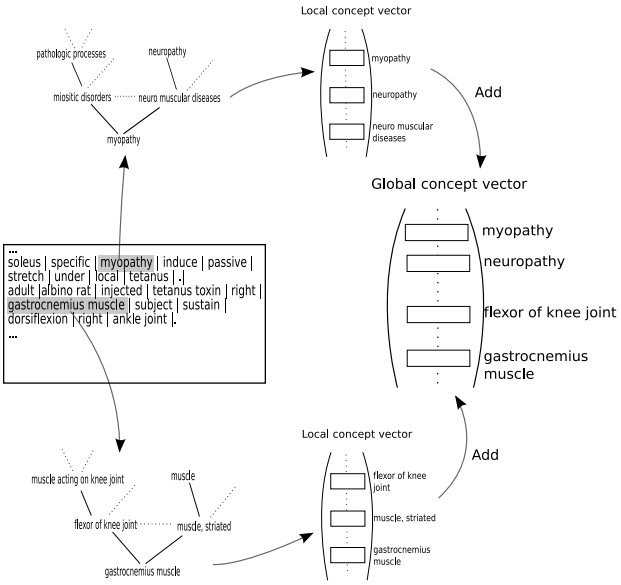

FIG. 1 – *La représentation d'un document texte selon le modèle d'espace vectoriel de concepts.*

comme des concepts abstraits de haut niveau. Il peut être intéressant d'analyser expérimentalement l'effet du mélange entre concepts linguistiques et concepts statistiques.
Nous définissons la matrice M de concept par document dans le CVSM comme ceci :

$$M = \begin{bmatrix} \frac{\Phi_1^{IDF}(d_1)}{||\mathbf{\Phi}^{IDF}(d_1)||} & \cdots & \frac{\Phi_1^{IDF}(d_n)}{||\mathbf{\Phi}^{IDF}(d_n)||} \\ \vdots & \ddots & \vdots \\ \frac{\Phi_m^{IDF}(d_1)}{||\mathbf{\Phi}^{IDF}(d_1)||} & \cdots & \frac{\Phi_m^{IDF}(d_n)}{||\mathbf{\Phi}^{IDF}(d_n)||} \end{bmatrix} \tag{7}$$

avec m la dimension du CVSM (le nombre total de concepts) et n le nombre de documents. La LSA étant une méthode non-supervisée, à savoir que la méthode n'utilise pas l'information sur les étiquettes des documents, nous utiliserons les documents étiquetés (documents d'apprentissage) et les documents non-étiquetés (documents de test) pour M.
La SVD de M donne :

$$M = U\Sigma V^T \tag{8}$$

Nous dénotons par U_k, les k vecteurs singuliers associés aux k valeurs singulières les plus élevées et Σ_k la matrice diagonale contenant les valeurs singulières.
La projection du vecteur $\mathbf{\Phi}^{IDF}(d)$ du CVSM dans l'espace sémantique latente est donnée par :

$$\mathbf{\Phi}^{lsa}(d) = \frac{\mathbf{\Phi}^{IDF}(d)}{||\mathbf{\Phi}^{IDF}(d)||}.U_k.\Sigma_k^{\frac{1}{2}} \tag{9}$$

Le noyau CVSM latent est alors défini par :

$$k_{LCVSM}(d_1, d_2) = \frac{\langle \mathbf{\Phi}^{lsa}(d_1), \mathbf{\Phi}^{lsa}(d_2) \rangle}{||\mathbf{\Phi}^{lsa}(d_1)||.||\mathbf{\Phi}^{lsa}(d_2)||} \tag{10}$$

5 Évaluation expérimentale

Nous avons mené plusieurs expérimentations sur un corpus médical pour évaluer la représentation CVSM. Nous avons utilisé le modèle d'espace vectoriel standard, VSM, comme base de comparaison. Dans cette section, nous présentons le corpus médical utilisé, le mode opératoire pour les expérimentations et un résumé des résultats expérimentaux.

5.1 Le corpus Ohsumed

Les expérimentations ont été menées sur le corpus Ohsumed qui est un corpus médical contenant 6286 documents d'apprentissage et 7643 documents de test (Hersh et al., 1994). Les documents sont des résumés d'articles de médecine issus de la base bibliographique médicale MEDLINE. Chaque document est, ou doit être dans le cas des documents de test, étiqueté avec une ou plusieurs des 23 étiquettes. Les étiquettes correspondent à des catégories cardio-vasculaires.
La tâche de catégorisation sur ce corpus est connue pour être difficile. En effet, les systèmes de catégorisation qui fonctionnent relativement bien sur les corpus tels que le Reuters-21578 et le 20-NewsGroups voient leurs performances diminuées. Par exemple, dans (Joachims, 1998), le classifieur linéaire SVM sur la VSM atteint une performance de 65.9% alors qu'il atteint une performance de 86% sur le corpus Reuters 21578. Les difficultés de catégorisation sont dues au fait que les données sont bruitées avec des termes médicaux très spécifiques et que les catégories ont un haut degré de corrélation.

5.2 La préparation des expérimentations

Dans toutes nos expérimentations, les documents ont été pré-traités pour la représentation selon le modèle d'espace vectoriel de concepts (CVSM). Pour le modèle d'espace vectoriel, VSM, nous avons utilisé le *stemming* pour réduire les mots fléchis à leurs bases communes. Nous avons, en outre, éliminés les mots vides. Pour le *stemming*, nous avons utilisé l'algorithme de Porter (Porter, 1980).

Pour la gestion du problème à multi-catégories, nous avons utilisé la stratégie "un-contre-tous". Cette stratégie a mené à la décomposition du problème principal en 23 sous-problèmes de catégorisation binaire. La librairie libSVM (Chang et Lin, 2001) a été utilisée pour l'apprentissage des classifieurs SVM.

Afin d'évaluer le CVSM, nous avons utilisé le VSM comme base de comparaison. Nous avons utilisé un noyau linéaire, dans le VSM, avec une pondération TF-IDF et une normalisation des vecteurs selon le mode opératoire defini dans (Joachims, 1998). Ce noyau est nommé "noyau Sac-de-mots"(*Bag Of Words - BOW - kernel*). De plus, nous avons aussi utilisé un

noyau LSA dans le VSM en utilisant l'équation 10. Ce noyau LSA est utilisé comme base de comparaison pour le noyau CVSM latent. Nous utiliserons la dénomination "*BOW SVD*" pour désigner le noyau LSA dans le VSM.

La mesure utilisée pour évaluer les performances des classifieurs est, principalement, la mesure F1 (Sebastiani, 2002). La mesure F1 est la moyenne harmonique de la précision et du rappel d'un système de catégorisation.

5.3 Évaluation de la puissance de décroissance α

Dans la première expérimentation, nous cherchons, empiriquement, la valeur optimale de la puissance de décroissance α pour le corpus Ohsumed. Nous rappelons que α contrôle la manière dont les concepts généraux sont pris en compte dans l'équation 3. Une valeur de zéro donnera un poids égal aux concepts généraux et spécifiques. Pour cette expérimentation, nous utilisons uniquement 10% des données d'apprentissage et 10% des données de test pour évaluer la performance. De plus, nous utilisons un échantillonnage stratifié pour conserver les proportions. La figure 2 montre les scores micro-F1 pour différentes valeurs de $\alpha \in [0, 1]$. Le meilleur score est obtenu pour $\alpha = 0.1$ avec une valeur micro-F1 de 53.9%. En outre, une meilleure performance est obtenue pour $\alpha = 0$ que pour $\alpha = 1$. Ce point signifie que les concepts généraux jouent un rôle important dans la tâche de catégorisation pour le corpus Ohsumed. C'est, effectivement, le cas lorsque plusieurs mots, avec un sens general commun, sont utilisés dans différents documents d'une même catégorie.

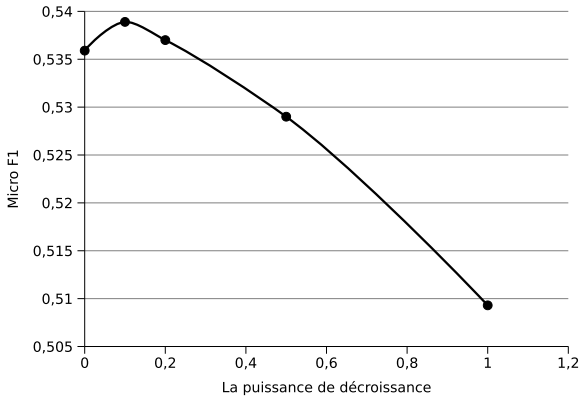

FIG. 2 – *La variation de la micro-F1 en fonction du pouvoir de décroissance α. Les résultats sont obtenus en utilisant 10% des données d'apprentissage et 10% des données de test.*

5.4 Évaluation du nombre de concepts latents

Pour les noyaux basés sur la LSA, la dimension de l'espace sémantique doit être fixée empiriquement. Par conséquent, nous avons mené un ensemble d'expérimentations sur le corpus

entier en faisant varier le nombre de concepts latents, à savoir le nombre de vecteurs singuliers associés aux valeurs singulières les plus élevées. La figure 3 montre la variation du score micro-F1. Les noyaux CVSM latent et BOW SVD atteignent leurs performances quasi-optimales pour approximativement 2000 concepts latents. En fait, il y a une croissance rapide des performances pour un nombre de concepts compris allant de 0 à 1000. Ceci montre que les 1000 premiers vecteurs singuliers fournissent l'information principale pour décrire les documents. Puis, la croissance des performances diminue pour un nombre de concepts de 1000 à 2000, indiquant ainsi la présence d'une faible quantité d'information.
Les performances du noyau CVSM latent sont meilleures de près de 2% par rapport au noyau BOW SVD. De plus, les différences sont plus prononcées pour un nombre de dimensions faible. Ceci signifie que le noyau CVSM latent est capable de capturer et d'exprimer l'information principale dans un espace de faible dimension, c'est à dire que l'information principale est résumée par un faible nombre de vecteurs singuliers.
Pour le reste des expérimentations, nous avons fixés le nombre de concepts latents à 3000.

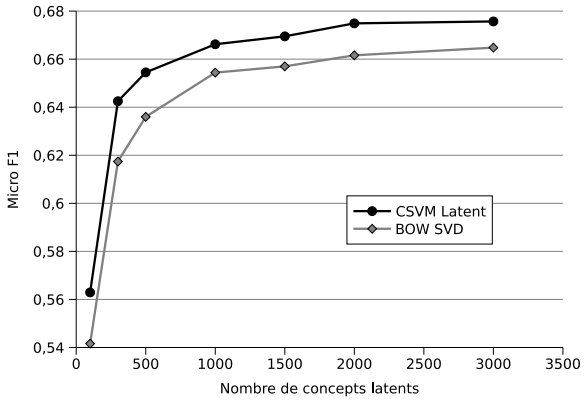

FIG. 3 – *La variation de la micro-F1 en fonction du nombre de concepts latents.*

5.5 Évaluation des noyaux

Les performances des noyaux sur le corpus Ohsumed sont reportées dans les tableaux 1 et 2. Les expérimentations ont été réalisées sur l'ensemble des données sans sélection de termes. Les noyaux dans le CVSM obtiennent de meilleurs résultats, jusqu'à 2% de plus, que les noyaux dans le VSM. Comme attendu, les noyaux LSA ont de meilleurs performances que les noyaux linéaires. Ceci est dû au fait que les concepts abstraits de haut niveau de l'espace sémantique résument l'information principale et réduisent le bruit. En outre, le noyau CVSM linéaire est plus performant que le noyau BOW SVD. Nous en déduisons que les concepts basés sur l'ontologie (les concepts linguistiques) sont plus adaptés pour exprimer le sens des documents médicaux que les concepts abstraits (les concepts statistiques) obtenus par la décomposition linéaire de la LSA.

Noyau	F1	Précision	Rappel
BOW Linéaire	65.83%	75.65%	58.26%
BOW SVD	66.48%	76.21%	58.95%
CVSM Linéaire	67.08%	77.77%	58.98%
CVSM Latent	67.57%	76.01%	60.82%

TAB. 1 – *Les scores micro-moyennés pour le corpus Ohsumed.*

Noyau	F1	Précision	Rappel
BOW Linéaire	60.32%	76.01%	51.7%
BOW SVD	60.62%	74.87%	52.69%
CVSM Linéaire	61.31%	78.07%	52.39%
CVSM Latent	62.71%	75.48%	55.29%

TAB. 2 – *Les scores macro-moyennés pour le corpus Ohsumed.*

5.6 Réduction de la quantité de données d'apprentissage

La figure 4 montre l'impact de la quantité de données utilisées pour l'apprentissage sur les performances des noyaux. Comme précédemment, les noyaux LSA sont plus performants que les noyaux linéaires. Néanmoins, il existe deux points intéressants dans ces résultats. Premièrement, le noyau BOW SVD est plus performant que le noyau CVSM linéaire lorsque 30% ou moins des données d'apprentissage sont utilisées. Au delà des 30% le noyau CVSM linéaire est plus performant. Ceci signifie que les concepts statistiques ont un pouvoir discriminatoire supérieur au concepts du CVSM pour des petits échantillons de données d'apprentissage. Toutefois, quand la base d'apprentissage est suffisamment importante, les concepts CVSM deviennent plus expressifs. Deuxièmement, tous les noyaux ne réussissent pas à capturer l'information principale avec une faible quantité de données. En effet, les performances ne cessent de s'améliorer en fonction de la quantité de données d'apprentissage. Par conséquent, nous pouvons en déduire que chaque document d'apprentissage fournit une nouvelle information qui améliore la performance de catégorisation. Ceci est principalement dû au fait que ce corpus contient des termes spécifiques avec une faible fréquence d'occurrence.

6 Conclusion

Dans cet article, nous avons présenté un modèle d'espace vectoriel de concepts pour la représentation de documents. Ce modèle utilise des connaissances a priori pour capturer le sens des documents textuels. Nous avons montré une façon simple d'utiliser efficacement les ontologies pour les intégrer au modèle d'espace vectoriel, VSM, standard. Nous avons, aussi, adaptés le noyau linéaire et le noyau LSA pour le CVSM. Nous avons illustré l'utilisation du CVSM par une application au domaine biomédical. Le *Metathesaurus* de l'UMLS a été utilisé comme source a priori de connaissance pour définir le CVSM.

Les performances des noyaux CVSM ont été, expérimentalement, évaluées sur une tâche de catégorisation de documents biomédicaux. Les noyaux ont été comparés au noyau sac-de-mots (BOW) et au noyau LSA. Les résultats ont montré que les noyaux CVSM améliorent

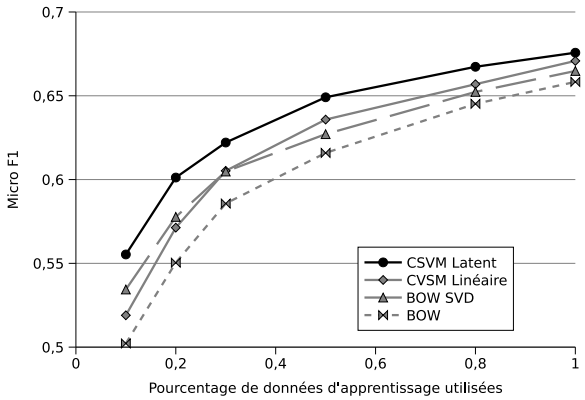

FIG. 4 – *La variation de la micro-F1 en fonction du pourcentage de documents d'apprentissage utilisés.*

les performances de catégorisation de près de 2%. De plus, les expérimentations ont montré que l'utilisation des concepts latents, extraits à partir des concepts linguistiques par une SVD, permettent d'améliorer les résultats.

Pour un travail futur, nous pensons améliorer la représentation CVSM en intégrant une pondération d'attributs plus adaptée que la pondération IDF. Il a été montré dans (Debole et Sebastiani, 2003) que la pondération supervisée des termes pouvait améliorer la catégorisation de documents. En effet, des recherches récentes, en matière de pondération des termes, ont donné des résultats prometteurs (Soucy et Mineau, 2005; Lan et al., 2006, 2007).

Références

Chang, C.-C. et C.-J. Lin (2001). *LIBSVM : a library for support vector machines*.

Cristianini, N., J. Shawe-Taylor, et H. Lodhi (2002). Latent Semantic Kernels. *Journal of Intelligent Information Systems 18*(2-3), 127–152.

Debole, F. et F. Sebastiani (2003). Supervised term weighting for automated text categorization. In *SAC '03 : Proceedings of the 2003 ACM symposium on Applied computing*, New York, NY, USA, pp. 784–788. ACM Press.

Deerwester, S. C., S. T. Dumais, T. K. Landauer, G. W. Furnas, et R. A. Harshman (1990). Indexing by Latent Semantic Analysis. *Journal of the American Society of Information Science 41*(6), 391–407.

Hersh, W., C. Buckley, T. J. Leone, et D. Hickam (1994). OHSUMED : an interactive retrieval evaluation and new large test collection for research. In *SIGIR '94 : Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*, New York, NY, USA, pp. 192–201. Springer-Verlag New York, Inc.

Joachims, T. (1998). Text categorization with support vector machines : learning with many relevant features. In *Proc. of ECML-98, 10th European Conference on Machine Learning*,

Heidelberg, DE, pp. 137–142. Springer Verlag.

Joachims, T. (2002). *Learning to Classify Text Using Support Vector Machines : Methods, Theory and Algorithms*. Norwell, MA, USA : Kluwer Academic Publishers.

Lan, M., C. L. Tan, et H. B. Low (2006). Proposing a new term weighting scheme for text categorization. In *AAAI'06 : Proceedings of the 21st National Conference on Artificial Intelligence*.

Lan, M., C. L. Tan, J. Su, et H. B. Low (2007). Text representations for text categorization : a case study in biomedical domain. In *IJCNN'07 : International Joint Conference on Neural Networks*.

Porter, M. F. (1980). An algorithm for suffix stripping. *Program 14*(3), 130–137.

Salton, G., A. Wong, et C. S. Yang (1975). A Vector Space Model for automatic indexing. *Communications of the ACM 18*(11), 613–620.

Sebastiani, F. (2002). Machine Learning in Automated Text Categorization. *ACM Computing Surveys 34*(1), 1–47.

Shawe-Taylor, J. et N. Cristianini (2004). *Kernel Methods for Pattern Analysis*. Cambridge University Press.

Soucy, P. et G. W. Mineau (2005). Beyond TFIDF Weighting for Text Categorization in the Vector Space Model. In L. P. Kaelbling et A. Saffiotti (Eds.), *IJCAI'05 : International Joint Conf. on Artificial Intelligence*, pp. 1130–1135. Professional Book Center.

Vapnik, V. N. (1995). *The nature of statistical learning theory*. New York, NY, USA : Springer-Verlag New York, Inc.

Summary

Kernels have been widely used in Natural Language Processing as similarity measures within inner-product based learning methods like Support Vector Machines. The Vector Space Model (VSM) has been extensively used for the spatial representation of the documents. However, it is purely a statistical representation. In this paper, we present a Concept Vector Space Model (CVSM) representation which uses linguistic prior knowledge to capture the meanings of the documents. We also propose a linear kernel and a latent kernel for this space. The linear kernel takes advantage of the linguistic concepts whereas the latent kernel combines statistical and linguistic concepts. Indeed, the latter kernel uses latent concepts extracted by the Latent Semantic Analysis (LSA) in the CVSM. The kernels were evaluated on a text categorization task in the biomedical domain. The Ohsumed corpus, well known for being difficult to categorize, was used. The results have shown that the CVSM improves the performances compared to the VSM.