S. Albitar, S. Fournier, B. Espinasse, « L'impact de l'enrichissement sémantique sur la classification de textes: Application au domaine médical », 6ème Atelier Recherche d'Information SEmantique RISE 2014, mar 2014 A paraitre

L'impact de l'enrichissement sémantique sur la classification de textes: Application au domaine médical

Shereen Albitar – Sébastien Fournier – Bernard Espinasse

Aix-Marseille Université, CNRS, LSIS UMR 7296, 13397, Marseille, France
{prénom.nom@lsis.org}

Résumé. L'utilisation de la sémantique dans la classification supervisée de texte peut améliorer son efficacité, en particulier, dans des domaines spécifiques. La plupart des travaux utilisent les concepts comme une alternative aux mots et transforment le classique sac de mots (BOW) en un sac de concepts (BOC). Cette transformation se fait à travers la tâche de conceptualisation. De plus, le BOC peut être enrichi par l'utilisation de concepts connexes par la prise en compte de ressources sémantiques pouvant ainsi améliorer l'efficacité de la classification. Cet article se focalise sur l'étude de l'impact, pour la classification supervisée de texte, de l'application d'une stratégie d'enrichissement sémantique à une représentation de texte déjà conceptualisée. Cette stratégie est basée sur une méthode d'enrichissement mutuel des vecteurs. Nous présentons une étude expérimentale pour évaluer cette stratégie d'enrichissement sémantique en utilisant la méthode de classification supervisée Rocchio dans le domaine médical, en utilisant l'ontologie UMLS (Unified Medical Language System) et le corpus Ohsumed. Grâ4ce à l'enrichissement sémantique, les résultats démontrent des améliorations significatives sur la classification de textes dans l'espace des concepts.

Mots-clés: classification supervisée de texte, sémantique, conceptualisation, enrichissement sémantique, mesures de similarité sémantique, domaine médical, UMLS, Rocchio

1 Introduction

La classification supervisée de textes est actuellement un sujet à la pointe de la recherche, en particulier dans des domaines tels que la recherche d'information, de la recommandation, de la personnalisation, des profils d'utilisateurs, etc. Parmi les méthodes les plus populaires pour la classification de texte, nous citons notamment la méthode Bayésienne (NB), les Machines à vecteurs de support (SVM) et Rocchio ou bien la classification basée sur les centroïds. Malgré leur popularité et les résultats corrects qu'elles affichent, ces méthodes, utilisant les sacs de mots (BOW) pour la représentation de texte, souffrent d'un manque de sémantique au niveau de la représentation de texte et ignorent tout aspect sémantique présent au sein du texte. Elles souffrent aussi d'un manque de sémantique au niveau du processus de classification lui-même. En outre, comme le montre [1], ces méthodes ont aussi des problèmes pour

gérer les classes larges (c'est-à-dire dont le spectre sémantique est étendu) et les classes peu peuplées (ayant peu d'exemples d'apprentissage). Ces méthodes ont aussi plus de difficultés à effectuer la tâche de classification lorsqu'elle est réalisée dans un domaine spécifique. Afin de résoudre ces différents types de problème, nous pensons que l'emploi de la sémantique semble être le plus approprié. De plus, de nombreux travaux montrent que l'utilisation de la sémantique dans la classification de texte peut améliorer son efficacité en particulier dans des domaines spécifiques [2, 3]. Dans le but d'utiliser la sémantique pour la classification supervisée de texte, plusieurs options s'offrent à nous. Il est possible d'utiliser la sémantique avant l'indexation, avant et après l'apprentissage et au moment de la prédiction de la classe. Toutefois, même si l'emploi de la sémantique semble prometteur, il nous semble important de mieux cerner dans quel cadre il est intéressant de l'employer, c'est-à-dire dans quel cas le gain est significatif par rapport aux méthodes classiques.

Dans ce travail, au travers un certain nombre d'expérimentations, nous essayons d'estimer l'impact d'une méthode d'enrichissement sémantique de la représentation sur la classification supervisée de texte. Il s'agit de la méthode *« Enriching Vectors »*. Dans ces travaux, nous avons choisi d'utiliser Rocchio [4], même s'il est moins performant que SVM pour certaines tâches, pour sa relative efficacité et sa simplicité en plus de son extensibilité par rapport à l'utilisation des ressources sémantiques dans le modèle d'apprentissage. En effet, Rocchio est capable d'utiliser la sémantique aussi bien dans sa représentation de texte au travers l'utilisation des sacs de concepts (BOC) qu'avant ou après l'apprentissage jusqu'à la phase de prédiction. Il est donc capable d'utiliser tout le spectre possible de l'implication de la sémantique dans la tâche de classification. Les expériences que nous comptons réaliser afin de tester cette méthode que nous présentons sont effectuées dans un domaine spécifique : le domaine médical. Pour cela, nous utilisons le corpus Ohsumed et la base de connaissances UMLS.

Dans la section 2, nous présentons un bref état des lieux des méthodes de classification de texte utilisant la sémantique. Ensuite, dans la section 3, nous présentons un cadre conceptuel général pour l'intégration de la sémantique dans le processus de la classification supervisée de texte en utilisant une stratégie d'enrichissement sémantique à partir d'une représentation BOC. Dans la section 4, nous présentons la stratégie d'enrichissement, basé sur l'enrichissement par la méthode des vecteurs enrichis « *Enriching Vectors »*. Dans la section 5, nous présentons brièvement Rocchio, les ressources sémantiques, le corpus Ohsumed, et les outils utilisés dans cette recherche. La section 6 présente notre processus d'expérimentation. Ensuite dans la section 7, nous présentons les résultats obtenus. Enfin, nous terminons par une évaluation de notre travail, suivi par différentes perspectives de recherche.

2 Classification supervisée de texte par usage de la sémantique

Selon la littérature, de nombreux travaux proposent des approches impliquant la sémantique dans la classification de texte à différents niveaux, par exemple, en faisant valoir l'utilité de la sémantique dans la représentation de texte [2, 5]. La plupart de ces

travaux ont transformé le classique sac de mots (BOW) représentant le texte dans l'espace vectoriel en sac de concepts (BOC) en choisissant les concepts comme une caractéristique alternative aux mots [3, 6]. Ils sont alors appliqués lors de l'indexation. D'autres travaux utilisent la similarité sémantique entre concepts ainsi que l'enrichissement de la représentation par les concepts. Ils sont généralement appliqués après l'indexation mais avant la prédiction. Trois grandes approches se distinguent pour l'enrichissement de la représentation du texte : (i) les noyaux sémantiques - généralement employés par les classifieurs SVM [2, 5, 7] , (ii) la généralisation [3], et (iii) l'enrichissement de vecteurs [6]. Cependant, les auteurs de [3] concluent que l'application de la généralisation pour les tâches de classification appliquées à un domaine spécifique provoque une détérioration de la performance. Enfin, il est possible d'impliquer la sémantique au niveau de la prise de décision en utilisant par exemple des mesures de similarité sémantique entre textes [8]. Dans cet article, notre travail se focalise plus particulièrement sur l'enrichissement de vecteurs.

3 Un cadre conceptuel pour la classification par enrichissement sémantique

La **Fig. 1** propose un cadre conceptuel résumant l'approche présentée dans cet article, impliquant la sémantique dans le processus de classification supervisée de textes. Dans cette approche, la sémantique est impliquée dans les différentes étapes du processus de classification : En premier, elle est impliquée lors de l'indexation au travers de la conceptualisation, puis en appliquant l'utilisation de la sémantique après l'apprentissage. La conceptualisation est le processus de recherche et de correspondance d'un concept pertinent provenant d'une ressource sémantique et qui traduit le sens d'un ou plusieurs mots d'un texte. Les concepts couvrant un document texte composent alors le vecteur sémantique qui représente le document en tant que BOC. L'utilisation de la sémantique après l'apprentissage se fait grâce à un enrichissement sémantique par l'usage de mesures de similarité sémantique.

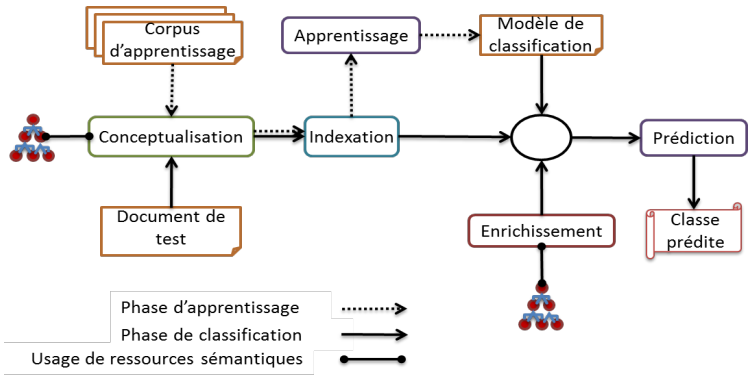

Fig. 1. Un framework conceptuel pour l'intégration de la sémantique dans la classification supervisée de textes

Dans ce travail, nous avons l'intention d'investiguer cette stratégie d'enrichissement qu'est « *Enriching Vectors* » en l'appliquant au domaine médical afin d'évaluer leur influence sur la classification supervisée de texte. L'étape de conceptualisation du texte est réalisée grâce à des travaux présentés dans [1, 9]. Ainsi, nous impliquons les connaissances sémantiques au niveau de l'indexation par l'utilisation de concepts au niveau de la représentation même du texte.

4 La méthode « Enriching Vectors »

Les auteurs dans [6] ont proposé cette méthode et l'ont appliquée pour de la catégorisation en utilisant K-Means et pour de la classification en utilisant kNN. Afin de comparer deux documents, les auteurs appliquent cette méthode sur les vecteurs représentant ces documents et ensuite applique une mesure de similarité classique comme Cosinus pour prédire la classe. Selon les auteurs, cette méthode a montré une meilleure corrélation avec un jugement humain, par rapport à l'application de la mesure de similarité classique sur les vecteurs d'origine sans enrichissement.

Les mesures de similarité classiques habituellement déployées pour comparer des documents de texte représentés dans l'espace vectoriel comme Cosinus dépendent d'une correspondance lexicale. En fait, ces mesures tiennent principalement compte des caractéristiques communes entre les vecteurs négligeant d'autres similitudes telles que la similarité sémantique entre des caractéristiques non partagées. En d'autres termes, si deux textes ne partagent pas les mêmes mots mais utilisent des synonymes, ils sont présumés dissemblables. Cet inconvénient a, entre autres, été souligné par [1].

Pour aller au-delà de la correspondance lexicale, nous avons l'intention d'appliquer *« Enriching Vectors »* à chaque paire de vecteurs avant la comparaison : chacun des vecteurs enrichit l'autre vecteur en utilisant ses caractéristiques exclusives. Étant donné deux documents A, B représentés à l'aide d'un vocabulaire de plusieurs concepts. Nous notons qu'une caractéristique est exclusive pour B, si elle est en correspondance avec un ou plusieurs mots du document B uniquement (la caractéristique n'est pas présente dans le document A) et réciproquement. Comme le montre la **Fig. 2**, l'objectif principal de cette approche est d'introduire les caractéristiques exclusives de B (C_2) dans A et de leur attribuer des poids appropriés en tenant compte des caractéristiques de A et vice versa. Ces pondérations sont estimées en utilisant les pondérations des autres caractéristiques du document traité et en utilisant la similarité sémantique entre ces caractéristiques et la caractéristique manquante. Pour ce faire, nous utilisons une matrice de proximité sémantique composée des similarités sémantiques entre les concepts du vocabulaire pair-à-pair.

Les nouveaux poids des concepts dans les vecteurs enrichis sont calculés comme suit :

$$w(c,A) = w\big(SC(c,A)\big) * sim\big(c, SC(c,A)\big) * CC(c,A)$$

Où $w\big(SC(c,\mathrm{A})\big)$ est le poids de la plus forte connexion du concept c (Strongest Connection) dans A ce qui correspond au poids du concept le plus similaire à c.

$sim\big(c, SC(c,\mathrm{A})\big)$ est la mesure de similarité entre c et le concept ayant la plus forte connexion dans A.

$CC(c, \mathrm{A})$ est la centralité contextuelle (CC) du concept c dans le document A qui est donné par la formule suivante :

$$CC(c, A) = \frac{\sum_{c_i \in A} sim(c, c_i) * w(c_i, A)}{\sum_{c_i \in A} w(c_i, A)}$$

Où :
$sim(c, c_i)$ est la similarité sémantique entre le concept c et c_i du document A.
$w(c_i, \mathrm{A})$ est le poids du concept c_i dans le document A.

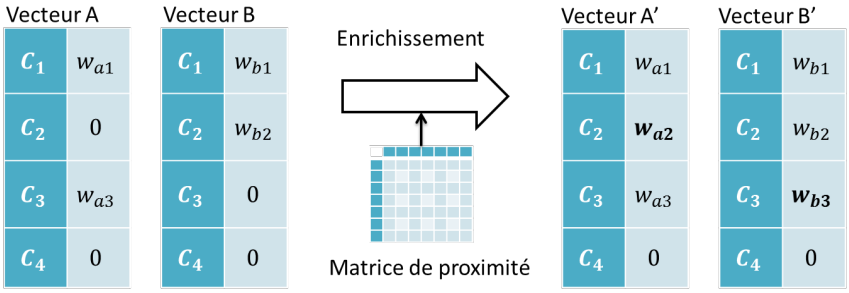

Fig. 2. Exemple d'enrichissement de vecteurs par la méthode *« Enriching Vectors »*

5 Les ressources et outils utilisés

Dans ces travaux, nous avons besoin des ressources sémantiques et de plusieurs outils. Ces derniers sont présentés dans la suite de cette section.

Nous utilisons dans nos travaux ***Rocchio*** pour la classification supervisée de textes. Dans Rocchio [4], chaque classe est représentée par un vecteur centroïd. Les centroïds obtenus lors de l'apprentissage représentent un modèle de classification qui résume les caractéristiques des documents de chaque classe. Au cours de la phase de classification, chaque document de test est comparé aux centroïds en utilisant des mesures de similarité non sémantique afin de lui attribuer la classe dont le centroïd est le plus similaire. Il existe un grand nombre de ces mesures de similarité, dans nos travaux nous en utilisons cinq : Cosinus, Jaccard, Kullback-Leibler, Levenshtein et Pearson [10]. L'utilisation de cinq mesures de similarité, au lieu d'utiliser seulement la plus connue : cosinus, nous permet d'estimer la différence d'impact de l'utilisation de la sémantique en fonction de la mesure de similarité et si cette différence est significative ou pas. La similarité, entre deux vecteurs A et B, est calculée selon les formules suivantes :

$$\mathrm{Sim}_{Cosinus}(\mathrm{A}, \mathrm{B}) = \cos(\mathrm{t}) = \frac{\sum_i a_i * b_i}{\sqrt{\sum_i a_i^2} * \sqrt{\sum_i b_i^2}}$$

$$\mathrm{Sim}_{Jaccard}(\mathrm{A}, \mathrm{B}) = \frac{\sum_i a_i * b_i}{\sqrt{\sum_i a_i^2} + \sqrt{\sum_i b_i^2} - \sum_i a_i * b_i}$$

$$Sim_{Levenshtein}(A, B) = 1 - (\sum |a_i - b_i| \,/\, \sum Max(a_i, b_i))$$

$$\mathrm{Sim}_{Pearson} = \frac{n\sum a_i b_i - \sum a_i \sum b_i}{\sqrt{\left[n\sum a_i^2 - (\sum a_i)^2\right]\left[n\sum b_i^2 - (\sum b_i)^2\right]}}$$

$$\mathrm{Sim}_{KullbackLeibler}(A,B) = \sum_i \left(\pi_1 * D(a_i||w_i) + \pi_2 * D(b_i||w_i)\right)$$

Où :

$\pi_1 = \frac{a_i}{a_i + b_i}$	$\pi_2 = \frac{b_i}{a_i + b_i}$
$w_i = \pi_1 * a_i + \pi_2 * b_i$	$D(a_i\|\|w_i) = a_i * \log(\frac{a_i}{w_i})$

Le corpus ***Ohsumed*** [11], utilisé pour l'apprentissage et les tests, est composé de résumés d'articles biomédicaux de l'année 1991 extraits de la base de données MEDLINE et indexés à l'aide de MeSH (Medical Subject Headings). Les premiers 20000 documents de cette base de données ont été sélectionnés et classés en utilisant 23 sous-concepts du concept « Disease ». Le corpus est alors divisé en deux parties : une pour l'apprentissage et l'autre pour les jeux d'essai. Dans ce travail, les centroïds des classes sont calculés par Rocchio pour chacune des cinq classes les plus fréquentes d'Ohsumed énumérées dans la **Table 1**.

Category	Description	Training	Test
C04	Neoplasms	972	1251
C06	Digestive System Diseases	588	632
C14	Cardiovascular Diseases	1192	1256
C20	Immune System Diseases	502	664
C23	Pathological Conditions, Signs and Symptoms	976	1181
Total		**4230**	**4984**

Table 1. Le Corpus Ohsumed

Unified Medical Language System (UMLS ®) [12] a été développé afin de modéliser le langage biomédical et celui de la santé. UMLS organise les concepts de différentes sources de vocabulaires (comme MeSH, SNOMED-CT, etc.) selon leurs sens en regroupant des concepts communs. Nous avons choisi, notamment pour des raisons de performance, d'effectuer la conceptualisation de textes en utilisant les concepts de SNOMED-CT uniquement.

En complément des ressources sémantiques comme UMLS, de nombreux outils ont été conçus afin de faciliter l'utilisation de ressources sémantiques pour le développement de systèmes médicaux. Dans ce travail, nous utilisons ***MetaMap*** [13] qui permet de faire la correspondance entre le texte et les concepts présents dans UMLS (et donc aussi dans SNOMED-CT).

L'outil ***UMLS-Similarity*** est un module Perl qui évalue la similarité sémantique entre les concepts d'UMLS. Nous utilisons dans ce travail cinq mesures de similarités sémantique issues de la version UMLS-similarity 1.33. Ces cinq mesures sont basées sur la structure de l'ontologie. Leur simplicité est à l'origine de leur efficacité qui a été

démontrée dans de nombreux domaines d'application dans lesquels les mesures de similarité sémantique sont utilisées [14]. Il s'agit de :

- *cdist* : cette mesure compte le nombre d'arêtes entre les concepts [15]. Son domaine de définition est compris entre zéro et deux fois la profondeur de l'ontologie. L'équation est la suivante :

  $$Sim_{Rada}(c_1, c_2) = min_{i\in[1,N]} |path_i(c_1, c_2)|$$

  Où :
  $path_i$ est le nombre de nœuds entre c_1 et c_2
  i est dans le domaine [1,N], N est le nombre de chemins possibles entre ces concepts dans l'ontologie.

- *wup* : cette mesure est calculée par deux fois la profondeur de la généralisation commune la plus spécifique des concepts (msca), divisé par la somme des profondeurs des concepts [16]. Son domaine de définition se situe entre zéro et un.

  $$Sim_{w\&P}(c_1, c_2) = \frac{2*depth(c)}{depth(c_1)+depth(c_2)} = \frac{2H}{N1+N2+2H}$$

  Où
  $N1$ et $N2$ correspondent aux nombres de connexions IS-A entre le concept commun le plus spécifique et c_1 et c_2 respectivement.
  H est le nombre de liens IS-A entre c et la racine de l'ontologie.

- *lch* : Cette mesure est le logarithme négatif du plus court chemin entre deux concepts divisé par deux fois la profondeur totale de l'ontologie [17]. Son domaine de définition va de 0 à la profondeur de l'ontologie.

  $$Sim_{lch}(c_1, c_2) = \text{Max}\left[-\log\left(\frac{Sim_{Rad}}{2D}\right)\right]$$

  Où : D est la profondeur maximum de l'ontologie et Sim_{Rad} est la similarité *cdist*.

- *zhong* : Cette mesure est la somme de la différence entre la « *milestone* » du msca et celle de chacun des concepts [18]. La « *milestone* » est un facteur calculé et est liée à la spécificité des concepts. Sa gamme se situe entre zéro et un.

  $$\text{milestone(c)} = \frac{1/2}{k^{depth(c)}}$$

  Où :
  Depth(c) est la profondeur du nœud c dans la hiérarchie
  k est une constance généralement de valeur 2.
  La distance est alors calculée comme suit :

  $$dc(c1, c2) = dc(c1, msca(c1, c2)) + dc(c2, msca(c1, c2))$$

  Où : $msca(c1, c2)$ est le plus proche parent commun de c_1 et c_2

$$dc(c, msca) = milestone(msca) - milestone(c)$$

- *nam* : c'est le logarithme d'une formule du chemin le plus court entre les deux concepts, et la profondeur de la taxonomie moins la profondeur du concept *msca* [19]. Son domaine dépend de la profondeur de la taxonomie.

$$Sim_{nam}(c_1, c_2) = log\left((Sim_{Rada} - 1)^{\alpha} * \left(Dc - depth(msca(c_1, c_2))\right)^{\beta} + K\right)$$

Pour des raisons de performance, nous avons introduit un nouvel outil : la matrice de proximité sémantique (**Fig. 2**). La matrice de proximité sémantique est une matrice carrée, dans laquelle chaque cellule correspond à la similarité sémantique entre chaque paire de concepts qui se trouvent dans l'index construit à partir de l'ensemble des documents. Ainsi, les similarités sémantiques sont utilisées au travers cette matrice de proximité sémantique dans le but d'enrichir, mutuellement, les représentations vectorielles.

6 Processus d'expérimentation

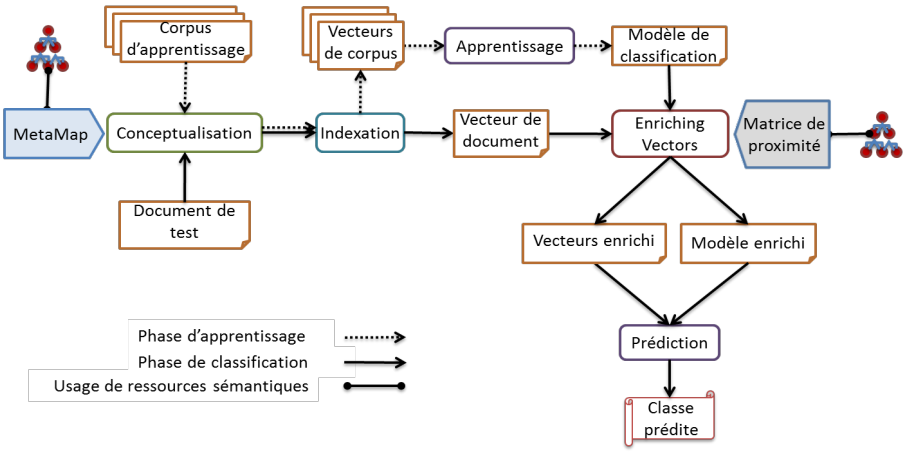

Fig. 3. L'enrichissement sémantique en utilisant « *Enriching Vectors* »

Afin d'évaluer l'effet de la méthode « *Enriching Vectors* » sur le processus de classification de texte à l'aide de Rocchio, nous utilisons la plate-forme expérimentale illustrée par la **Fig. 3**. Cette plate-forme utilise Rocchio pour l'apprentissage et pour la prédiction. L'étape de la conceptualisation est réalisée en amont par l'outil MetaMap avant d'effectuer l'étape d'indexation. Ainsi, les mots dans les documents du corpus sont entièrement substitués par les concepts trouvés par MetaMap ce qui permet d'indexer le corpus en tant que BOC. Pendant l'étape d'enrichissement, le vecteur du document de test est comparé à chacun des centroïds appris pendant l'apprentissage dans l'espace de concepts. Ils sont alors mutuellement enrichis en utilisant la matrice de proximité sémantique de l'une des cinq mesures de similarité sémantique. Après

cet enrichissement, les vecteurs traités sont moins espacés dans l'espace et partagent plus de caractéristiques communes (concepts). Enfin, l'étape de prédiction applique l'une des mesures de similarité classiques et les résultats sont ensuite évalués.

Dans ces expériences, la plate-forme exécute l'apprentissage une fois. Durant la classification, nous utilisons pour chaque expérimentation une des cinq mesures de similarités sémantiques pour l'enrichissement (cdist, lch, nam, wup, zhong) et une variante de Rocchio en utilisant une des cinq mesures de similarité classiques.

7 Résultats

Les résultats détaillés des exécutions qui sont liés à chaque mesure de similarité sémantique sont regroupés afin d'analyser l'impact d'« Enriching Vectors » sur l'efficacité des cinq variantes de Rocchio. Les résultats des cinq variantes sont illustrés par la **Fig. 4**. Les résultats de l'expérimentation conduisent à soulever les points suivants :

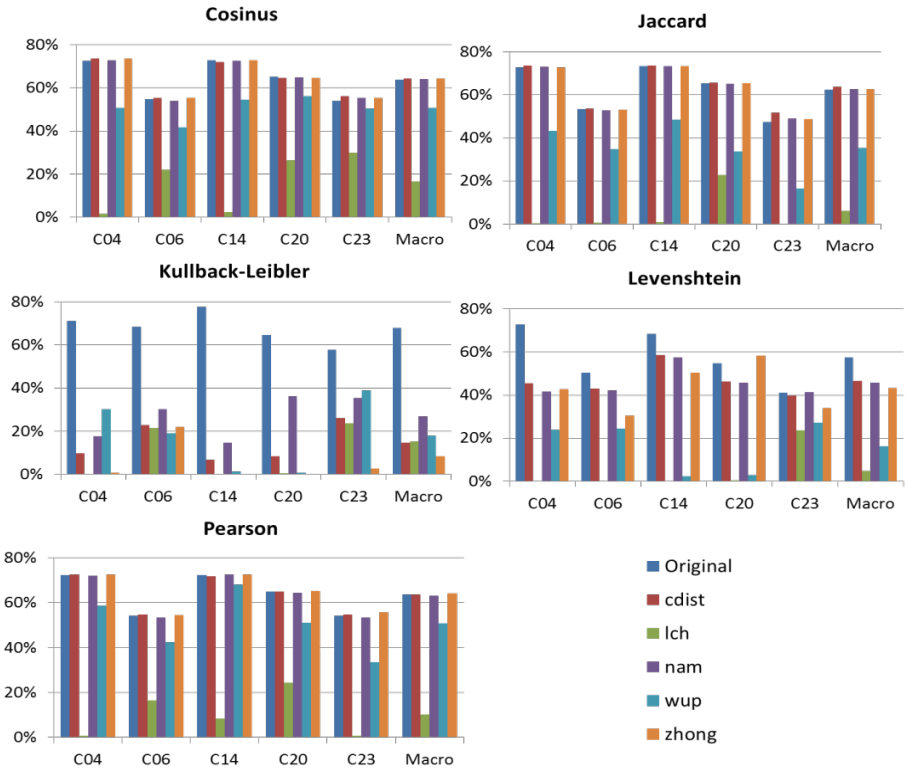

Fig. 4. F1-Measure avant et après l'application de la méthode « *Enriching Vectors* » en utilisant cinq mesures de similarité sémantique

Tout d'abord, dans tous les cas, l'utilisation des mesures de similarités lch et wup a causé une détérioration des performances de Rocchio tandis que les autres mesures de similarité ont montré des améliorations. Notons que le seul aspect que partagent cdist,

nam, et zhong est d'avoir un domaine de fonction (entre 0 et 1) par rapport à lch et wup ce qui peut justifier l'influence différente qu'ils peuvent avoir sur la représentation de texte. La meilleure performance globale a été obtenue à l'aide de la variante de Rocchio utilisant cosinus et zhong avec une macro moyenne de la F1-mesure de (64,33 %). Cette valeur est plus élevée que celle rapportée dans [6], qui est de (59,1%), où les auteurs ont testé « *Enriching Vectors* » sur un petit corpus extrait d' Ohsumed en utilisant le classifieur kNN.

Deuxièmement, on distingue deux groupes de variantes de Rocchio selon leur performance après l'application d'*« Enriching Vectors »* : le premier groupe contient Cosinus, Jaccard et Pearson et le second contient KullbackLeibler et Levenshtein. La principale différence entre ces deux groupes est que le premier évalue la similarité entre les vecteurs en utilisant leurs concepts communs tandis que le second dépend du nombre de leurs concepts discriminants afin d'évaluer leurs similarités. En général, « *Enriching Vectors »* vise à réduire la dispersion de la représentation des textes dans l'espace de concepts, ce qui semble aider le premier groupe dans l'évaluation des similarités. Au contraire, cet enrichissement semble être néfaste pour l'évaluation des similarités qui se basent sur les concepts discriminants entre les vecteurs.

Troisièmement, lorsqu'un système de classification classique obtient une faible valeur de la F1-mesure, *« Enriching Vectors »* a pu améliorer cette valeur. En effet, c'est le cas de la classe (C23) dont les résultats sont détaillés dans la **Table 2**. Le gain maximal a atteint (9,45%) dans le cas de la variante de Rocchio utilisant Jaccard après l'enrichissement des vecteurs en utilisant la mesure cdist. Ces résultats sont similaires à nos observations lors de l'application de la conceptualisation [9]. En fait, la classe « C23 » est sémantiquement très large par rapport aux autres classes et la représentation enrichie, par des concepts similaires, des documents et du modèle abouti à une meilleure identification de cette classe, ce qui a conduit à de meilleurs résultats.

Enrichissement	Cosinus		Jaccard		Kullback-Leibler		Levenshtein		Pearson	
BOC Original	**53,96**		**47,40**		**57,69**		**41,03**		**54,20**	
cdist	56,17	***+4,10****	51,88	***+9,45****	26,04	-54,86	39,69	-3,28	54,73	***+0,97****
lch	29,84	-44,69	0,00	-100,00	23,71	-58,89	23,50	-42,74	0,67	-98,76
nam	55,37	***+2,63****	49,16	***+3,71****	35,46	-38,52	41,32	***+0,69***	53,30	**-1,65**
wup	50,46	-6,48	16,61	-64,97	38,95	-32,48	27,15	-33,83	33,47	-38,25
zhong	55,26	***+2,41****	48,73	***+2,79****	2,58	-95,52	33,89	-17,41	55,73	***+2,82****

Table 2. Résultats de classification des documents de la classe C23. Les valeurs sont des F1-mesure (pourcentage). Les * signifient que les améliorations sont significatives selon McNemar

Enfin, il semble bénéfique à la classification basée sur Rocchio d'appliquer *« Enriching Vectors »* avant la prédiction car le comportement du classifieur semble être modifié et peut améliorer son efficacité. Cependant, le bénéfice obtenu est fonction de la mesure de similarité sémantique utilisée pour l'enrichissement et également de la mesure de similarité utilisée pour la prédiction. Par conséquent, il est nécessaire d'investiguer expérimentalement les bénéfices obtenus pour vérifier si *« Enriching Vectors »* est utile dans un contexte particulier.

8 Conclusion

A travers des expériences dans le domaine biomédical avec le corpus Ohsumed, l'ontologie UMLS, et la méthode de classification supervisée Rocchio, nous avons essayé d'estimer l'impact d'une stratégie d'enrichissement sémantique pour la classification supervisée de textes.

L'enrichissement, réalisé après l'entraînement et avant la prédiction des classes, est basée sur la méthode *« Enriching Vectors »*. Les résultats obtenus après l'enrichissement sont meilleurs de ceux obtenus sur les BOCs sans enrichissement pour plusieurs classes de documents. Cette amélioration est significative particulièrement pour la classe « C23 » qui est une classe large et hétérogène difficile à classer. Néanmoins, ces améliorations dépendent très largement de la mesure de similarité sémantique utilisée dans l'enrichissement et de la mesure de similarité utilisée pour la prédiction. Nous avons constaté également que l'enrichissement sémantique mutuel des vecteurs est bénéfique en utilisant les mesures de similarité qui se basent sur les caractéristiques communes entre les vecteurs comparés.

Dans de futurs travaux, nous avons l'intention de tester d'autres familles de mesures de similarité sémantique comme les mesures basées sur le contenu d'information (IC) ou basées sur les caractéristiques en les testant sur Ohsumed et sur d'autres corpus médicaux, comme celui de « TREC genomics » ou de « i2b2 ».

9 Références

[1] S. Albitar, S. Fournier, and B. Espinasse, "The Impact of Conceptualization on Text Classification," presented at the Proceedings of the 13th international conference on Web Information Systems Engineering, Paphos, Cyprus, 2012.

[2] S. Aseervatham and Y. Bennani, "Semi-structured document categorization with a semantic kernel," *Pattern Recogn.,* vol. 42, pp. 2067-2076, 2009.

[3] S. Bloehdorn and A. Hotho, "Boosting for text classification with semantic features," presented at the Proceedings of the 6th international conference on Knowledge Discovery on the Web: advances in Web Mining and Web Usage Analysis, Seattle, WA, 2006.

[4] E.-H. Han and G. Karypis, "Centroid-Based Document Classification: Analysis and Experimental Results," presented at the 4th European Conference on Principles of Data Mining and Knowledge Discovery, 2000.

[5] D. Ó. Séaghdha, "Semantic classification with WordNet kernels," presented at the Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers, Boulder, Colorado, 2009.

[6] L. Huang, D. Milne, E. Frank, and I. H. Witten, "Learning a concept-based document similarity measure," *J. Am. Soc. Inf. Sci. Technol.,* vol. 63, pp. 1593-1608, 2012.

[7] P. Wang and C. Domeniconi, "Building semantic kernels for text classification using wikipedia," in *14th ACM SIGKDD international conference on Knowledge discovery and data mining*, Las Vegas, Nevada, USA, 2008, pp. 713-721.

[8] M. Mohler and R. Mihalcea, "Text-to-text semantic similarity for automatic short answer grading," in *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, Athens, Greece, 2009, pp. 567-575.

[9] S. Albitar, S. Fournier, and B. Espinasse, "Conceptualization Effects on MEDLINE Documents Classification Using Rocchio Method," in *Web Intelligence*, ed, 2012, pp. 462-466.

[10] A. Huang, "Similarity measures for text document clustering," presented at the Sixth New Zealand Computer Science Research Student Conference, , Christchurch, New Zealand, 2008.

[11] W. Hersh, C. Buckley, T. J. Leone, and D. Hickam, "OHSUMED: an interactive retrieval evaluation and new large test collection for research," in *17th annual international ACM SIGIR conference on Research and development in information retrieval*, Dublin, Ireland, 1994, pp. 192-201.

[12] UMLS®. (2013). *Unified Medical Language System*. Available: http://www.nlm.nih.gov/research/umls/

[13] A. R. Aronson and F. M. Lang, "An overview of MetaMap: historical perspective and recent advances," *J Am Med Inform Assoc,* vol. 17, pp. 229-36, May-Jun 2010.

[14] T. Pedersen, S. V. S. Pakhomov, S. Patwardhan, and C. G. Chute, "Measures of semantic similarity and relatedness in the biomedical domain," *J. of Biomedical Informatics,* vol. 40, pp. 288-299, 2007.

[15] J. E. Caviedes and J. J. Cimino, "Towards the development of a conceptual distance metric for the UMLS," *J. of Biomedical Informatics,* vol. 37, pp. 77-85, 2004.

[16] Z. Wu and M. Palmer, "Verbs semantics and lexical selection," presented at the Proceedings of the 32nd annual meeting on Association for Computational Linguistics, Las Cruces, New Mexico, 1994.

[17] C. Leacock and M. Chodorow, "Combining Local Context and WordNet Similarity for Word Sense Identification," in *WordNet: An Electronic Lexical Database (Language, Speech, and Communication)*, C. Fellbaum, Ed., ed: The MIT Press, 1998, pp. 265-283.

[18] J. Zhong, H. Zhu, J. Li, and Y. Yu, "Conceptual Graph Matching for Semantic Search," presented at the Proceedings of the 10th International Conference on Conceptual Structures: Integration and Interfaces, 2002.

[19] H. Al-Mubaid and H. A. Nguyen, "A Cluster-Based Approach for Semantic Similarity in the Biomedical Domain," in *Engineering in Medicine and Biology Society, 2006. EMBS '06. 28th Annual International Conference of the IEEE*, 2006, pp. 2713-2717.