Chapitre 7 Filtrage de textes représentés par des "sacs de mots"

Ce chapitre expose les résultats obtenus sur les corpus Reuters et TREC-8 avec un réseau à une couche cachée et une représentation des textes en sac de mots. Afin de bien comprendre les résultats, l'accent est mis sur la sélection de descripteurs pour mieux mesurer l'influence des différents paramètres.

Ce chapitre se termine par la description de notre participation à la compétition TREC-8.

7.1 Présentation du modèle

Le premier modèle utilisé est un modèle simple, qui possède une architecture classique de perceptron multi-couche rappelée sur la Figure 7.1, avec des fonctions d'activation sigmoïdes pour les neurones cachés et une fonction d'activation logistique pour le neurone de sortie.

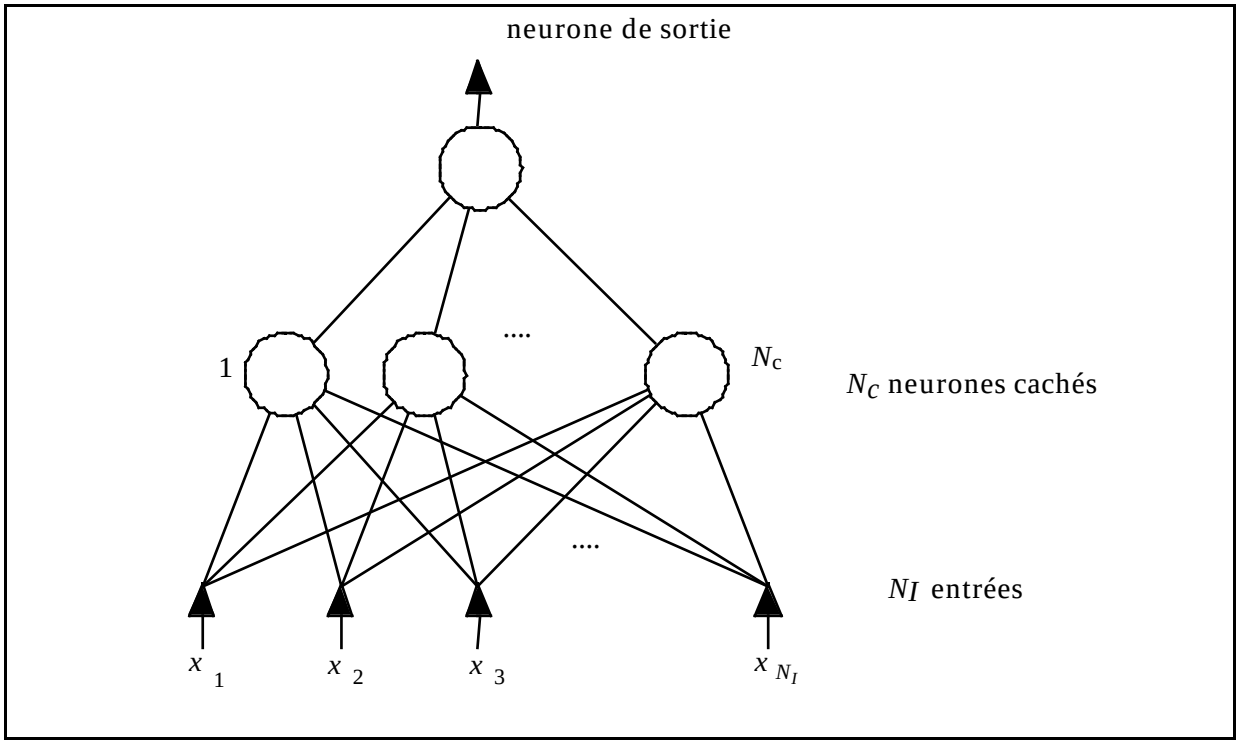

Figure 7.1 : *Architecture utilisée pour le modèle simple (les biais ne sont pas représentés pour plus de lisibilité). Les entrées x_i sont des fonctions des fréquences des mots auxquels sont associées les entrées.*

Dans ce modèle, les descripteurs utilisés pour représenter les textes sont uniquement des mots simples, sélectionnés par une méthode de sélection de descripteurs (cf. chapitre 5) ; le nombre de neurones cachés détermine la complexité de la fonction de classification.

Pour fabriquer un filtre, il est nécessaire :

- de définir les caractéristiques du vecteur des entrées : il faut déterminer sa dimension ainsi que la nature de ses composantes.
- de déterminer le nombre N_c de neurones cachés optimal afin d'obtenir les meilleures performances sur une base de test différente de la base d'apprentissage.
- de choisir une méthode de régularisation pendant la phase d'apprentissage comme indiqué dans le chapitre 6 : arrêt prématuré ou *weight decay* et, dans ce dernier cas, il faut déterminer les hyperparamètres intervenant dans la fonction de coût.

Chacun de ces choix a une grande influence sur la qualité du filtre obtenu ; nous allons décrire en détail la mise en œuvre de ces différentes étapes.

7.2 Sélection des descripteurs par la méthode de Gram-Schmidt

Le chapitre 5 a permis de montrer que, parmi les trois méthodes de sélection de descripteurs étudiées (vocabulaire spécifique, information mutuelle et orthogonalisation de Gram-Schmidt), aucune ne prévalait nettement sur les autres.

Néanmoins, la méthode d'orthogonalisation de Gram-Schmidt, couplée au critère d'arrêt décrit au chapitre 5, présente des avantages par rapport aux deux autres méthodes ; nous allons donc chercher à mieux comprendre son comportement.

Les différentes étapes de la mise en œuvre de la méthode d'orthogonalisation de Gram-Schmidt sont les suivantes :

1. À partir de l'ensemble des textes pertinents, on détermine une liste du vocabulaire spécifique du sous-ensemble des documents non pertinents.
2. À partir des textes pertinents, d'un sous-ensemble de textes non pertinents et de la liste du vocabulaire spécifique, on construit une matrice *X(N, Q)* où *N* est le nombre d'exemples total et *Q* est le nombre de descripteurs initiaux.

3. On met en œuvre l'algorithme d'orthogonalisation de Gram-Schmidt avec l'utilisation d'un vecteur aléatoire pour déterminer le critère d'arrêt.

La construction de la matrice *X* intermédiaire nécessite donc d'effectuer plusieurs choix :

1. Le codage des fréquences des descripteurs.
2. Le nombre *Q* de descripteurs initiaux.
3. Le nombre *N* d'exemples, et plus précisément le nombre d'exemples pertinents et le nombre d'exemples non pertinents.

La procédure de sélection de descripteurs détermine la qualité des descripteurs sélectionnés et leur nombre, cette qualité se mesurant grâce à la performance du filtre construit. Par conséquent, l'évaluation de la qualité de la sélection de descripteurs dépend du modèle construit à partir de cette sélection, mais elle ne peut pas être mesurée intrinsèquement. Afin de pouvoir comparer les méthodes de sélection de descripteurs, les techniques d'apprentissage (algorithmes, hyperparamètres) que nous mettrons en œuvre seront toujours identiques.

Ces choix sont tous interdépendants, et dépendent également de la nature du thème traité. Il est difficile de les étudier séparément, mais il existe des règles de conduite que nous allons mettre en évidence ci-après.

7.2.1 Codage des matrices

7.2.1.1 Centrage et normalisation des données

Le calcul intervenant dans la procédure de classement fait intervenir des produits scalaires, notamment le calcul du cosinus carré entre le vecteur de sortie et les vecteurs de descripteurs X_p :

$$\cos^2(X_p,Y) = \frac{\left[X_p^T.Y\right]^2}{\left[X_p^T.X_p\right].\left[Y^T.Y\right]}$$

Le codage de ces vecteurs influe sur ces produits scalaires, et par conséquent sur le classement trouvé.

Si l'on considère un ensemble d'exemples comportant T_1 exemples pertinents et T_0 exemples non pertinents, et que le vecteur de sortie représente par le nombre S_1 le premier ensemble et par S_0 le second, le produit scalaire entre un vecteur de descripteurs X_p et le vecteur de sortie Y s'écrit :

$$\left[\quad \right]^2 - \left(\quad \right)$$

Cette expression montre tout d'abord qu'il n'est pas possible de choisir une valeur nulle pour S_1 ou S_0 puisque cela reviendrait à ne pas tenir compte de l'une des deux classes.

Il est simple d'étudier certains cas limites. Par exemple, si un mot apparaît avec la même fréquence dans chaque texte quelle que soit sa classe, alors la valeur du cosinus doit être nulle puisque ce descripteur n'apporte aucune information. Si A est la valeur que prend ce descripteur, le produit scalaire s'écrit :

$$\left[X_p.Y\right]^2 = A^2\left(S_1.T_1 + S_0.T_0\right)^2 = 0$$

Comme cette relation doit être vraie quel que soit A, il est nécessaire d'avoir :

$$S_1.T_1 + S_0.T_0 = 0$$

Cette relation implique qu'il est nécessaire d'avoir un vecteur des sorties de moyenne nulle.

Dans la pratique, on utilise le codage suivant :

$$S_0 = \frac{-2.T_1}{T_1 + T_0} \text{ et } S_1 = \frac{2.T_0}{T_1 + T_0}$$

qui correspond à un codage initial de +1 pour les textes pertinents et –1 pour les textes non pertinents, puis à un centrage de ce vecteur.

De même, il est nécessaire de centrer la matrice X des descripteurs, car, dans le cas contraire, l'absence d'un descripteur est codée 0, ce qui est une valeur particulière dans la somme du produit scalaire.

7.2.1.2 Codage des fréquences

Le codage de la matrice X tient compte des fréquences de chaque terme dans les textes. Il provient du codage utilisé dans [Singhal, 1996] qui prend en considérations les variations de longueurs entre les textes : si $TF_j(i)$ est la fréquence du descripteur i dans le texte j, et si $\overline{TF_j}$ est la fréquence moyenne dans le texte j, on a :

$$x_j^i = \frac{1 + \log\left(TF_j(i)\right)}{1 + \log\left(\overline{TF_j}\right)}$$

7.2.2 Impact des caractéristiques de la matrice ***X***

7.2.2.1 Problématique

La méthode d'orthogonalisation de Gram-Schmidt nécessite la construction d'une matrice *X(N, Q)* constituée de *N* exemples, chacun étant décrit par *Q* descripteurs.

Pour chaque thème, l'ensemble des *N* exemples est divisé en deux sous-ensembles : N^+ est le nombre de documents pertinents et N^- le nombre de documents non pertinents.

$$N = N^+ + N^-$$

L'étude des corpus, présentée au chapitre 3, a montré que le nombre d'exemples pertinents pour un thème est en général limité : il semble raisonnable de prendre en considération l'ensemble des documents pertinents disponibles. En revanche, le nombre d'exemples non pertinents dont on dispose est, en général, très grand, puisque ce sont tous les documents du corpus non pertinents pour un thème.

Il est donc nécessaire de déterminer le nombre de textes non pertinents à prendre en considération, et de choisir ceux-ci.

- Pour le corpus Reuters, la base d'apprentissage disponible (documents datés avant le 8 avril 1987) comporte 9600 textes ; donc pour la catégorie *earn* qui comporte le plus de documents pertinents, il reste 6723 documents non pertinents possibles ; pour les catégories comportant le moins de documents pertinents, il reste 9599 documents disponibles.
- Pour le corpus de TREC-8, la base du *Financial Times 1992* qui est utilisée pour constituer les bases d'apprentissage comporte 64139 documents, il reste donc pour chaque thème environ 64000 textes utilisables pour former le sous-ensemble des documents non pertinents.
- Pour le corpus de l'AFP, le nombre de documents pertinents représente également plusieurs centaines de milliers de documents potentiels.

En conséquence, choisir le nombre *N* d'exemples revient à choisir l'ensemble des documents non pertinents.

Les *Q* descripteurs sont les *Q* premiers mots de la liste du vocabulaire spécifique du sous-ensemble des documents pertinents. Les mots sélectionnés par la méthode seront sélectionnés parmi ces *Q* descripteurs qui sont appelés dans la suite *descripteurs initiaux*.

Pour effectuer la procédure de sélection des descripteurs, il faut choisir à la fois le nombre *N* d'exemples et le nombre *Q* de descripteurs initiaux. L'expérience décrite ci-dessous étudie l'influence des choix de *N* et de *Q* sur les résultats de la sélection de descripteurs.

7.2.2.2 Description de l'expérience

Pour cette expérience, le modèle retenu est une simple régression logistique comme présenté à la Figure 7.2. La détermination du vecteur de poids *w* est effectuée en minimisant une fonction de coût *J'* incluant un terme de *weight decay* par les méthodes de minimisation décrites au chapitre 6. La fonction *J'* s'écrit :

$$J'(w) = EC(\ ) \quad -\|\ \|$$

EC(w) est l'entropie croisée et l'hyperparamètre α est fixé à 1.

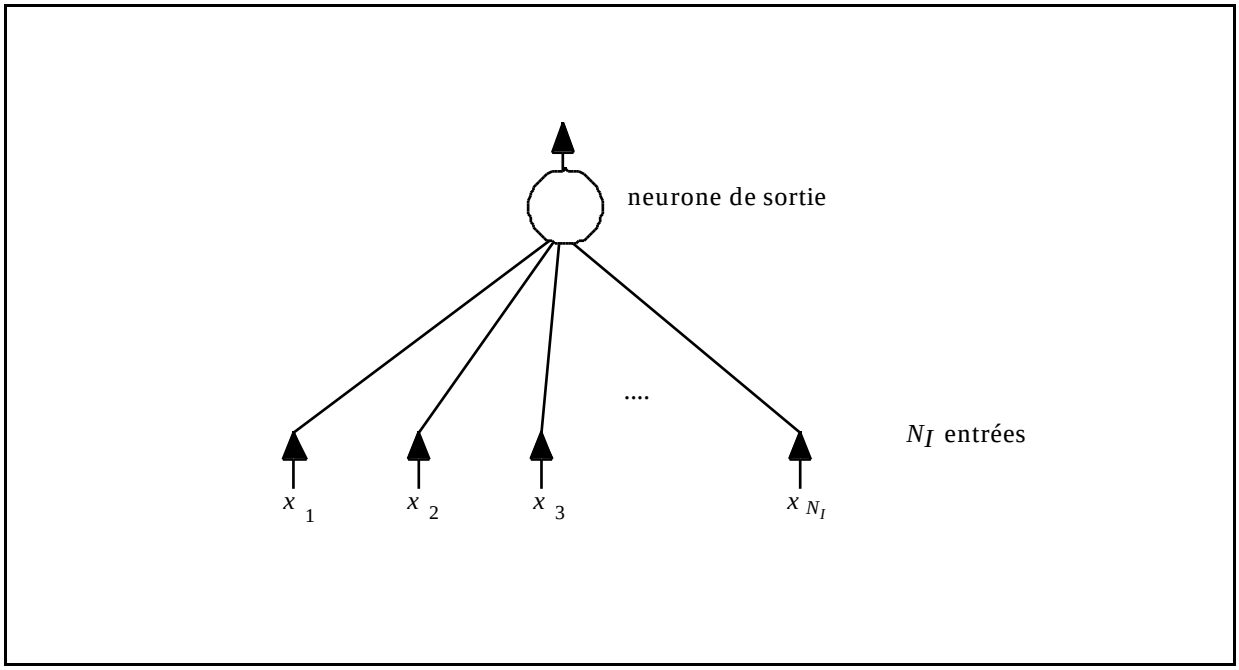

Figure 7.2 : *Régression logistique équivalente à un réseau avec 0 neurone caché (le biais n'est pas représenté).*

Les expériences sont menées sur 5 thèmes issus du corpus Reuters dont les caractéristiques sont rappelées Figure 7.3. Ces thèmes ont été choisis, d'une part, parce qu'ils ne sont pas trop faciles car ils ne dépendent pas exclusivement de la présence ou de l'absence d'un seul mot-clef. D'autre part, le nombre de documents pertinents est variable selon les thèmes, ce qui nous a permis d'étudier les corrélations éventuelles entre les valeurs des paramètres N et Q et le nombre de documents pertinents sur la base d'apprentissage.

Catégorie	Apprentissage	Test
interest	347	131
oilseed	124	47
nat-gas	75	30
sorghum	24	10
lumber	10	6

Figure 7.3 : *Liste des thèmes étudiés avec le nombre de documents pertinents disponibles sur chaque base.*

Pour chacun de ces thèmes, le nombre Q de descripteurs initiaux utilisés pour la sélection de descripteurs prend les valeurs 10, 50, 100, 200, 400.

Pour chacune des valeurs de Q, le nombre de documents non pertinents (donc le nombre N) varie ; les valeurs suivantes ont été testées : 100, 500, 1000, 2000, 3000, 4000, 5000.

7.2.2.3 Résultats des expériences

Pour chacune des expériences, les performances après l'apprentissage sont évaluées sur la base de test, par la valeur de F optimale (définie au chapitre 4) ou par les courbes rappel-précision interpolée. La valeur de F optimale est obtenue en testant plusieurs valeurs du seuil de décision pour la sortie du classifieur (de 0 à 1 par pas de 0.1) et en conservant la meilleure valeur de F sur la base de test pour s'affranchir du choix du seuil.

Les résultats obtenus pour chaque thème sont présentés sur la Figure 7.4. Pour un même thème, chaque courbe représente une valeur de Q différente, l'axe des abscisses représentant le nombre N^- de documents non pertinents utilisés pour la construction de la matrice X. Les performances sont présentées sur la colonne de gauche tandis que la colonne de droite précise le nombre de descripteurs sélectionnés par le critère d'arrêt pour chaque expérience.

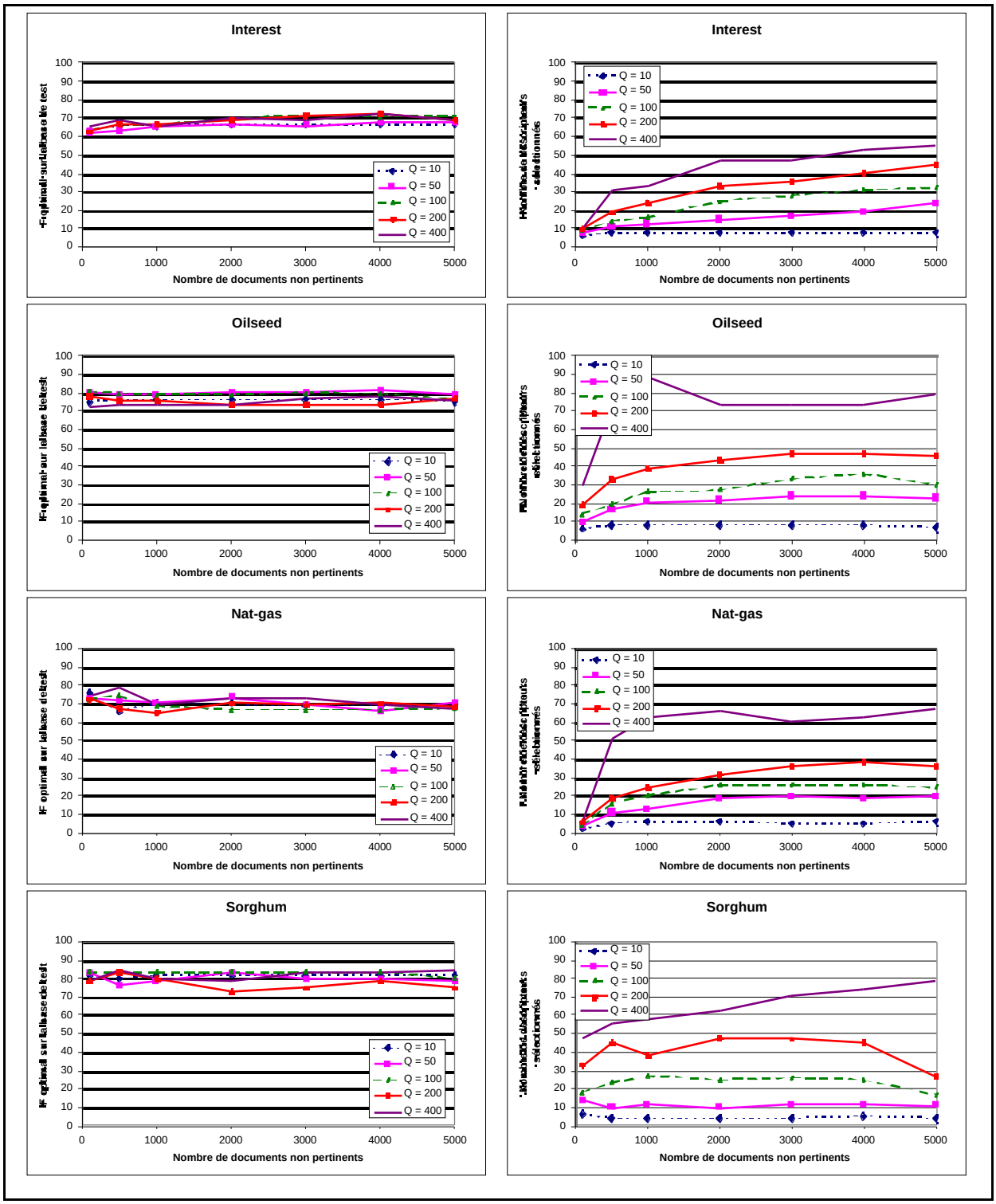

Figure 7.4 : *La colonne de gauche montre la valeur de F optimale pour chaque thème et la colonne de droite le nombre de descripteurs sélectionnés en fonction du choix de Q et du nombre de documents non pertinents.*

7.2.2.4 Interprétation des résultats

L'ensemble des résultats montre que le nombre initial de descripteurs et le nombre d'exemples non pertinents utilisés ont une influence sur les résultats de la sélection de descripteurs, et notamment sur le nombre de descripteurs sélectionnés.

Les trois premiers thèmes *interest*, *oilseed* et *nat-gas*, qui ont le plus de documents pertinents sur la base d'apprentissage, semblent avoir des comportements réguliers et de même nature tandis que les deux derniers thèmes *sorghum* et *lumber* ont un comportement différent des trois premiers.

Ces observations montrent que le choix des paramètres N et Q doit se faire en fonction du nombre de documents pertinents sur la base d'apprentissage.

Pour les trois premiers thèmes, on peut remarquer en ce qui concerne le nombre de descripteurs sélectionnés que :

1. Pour un nombre de descripteurs initial constant, le nombre de descripteurs finalement sélectionnés est d'autant plus grand que le nombre de documents non pertinents est élevé.
2. Pour un nombre de documents non pertinents fixés, le nombre de descripteurs sélectionnés est d'autant plus grand que le nombre de descripteurs initial est élevé.
3. Le facteur prédominant pour le nombre de descripteurs sélectionnés est le nombre initial de descripteurs.

Pour le dernier thème qui ne comporte que dix documents pertinents, les tendances sont différentes selon le nombre initial de descripteurs, les courbes correspondant à 200 et 400 descripteurs initiaux se détachant des autres courbes.

Les courbes de la colonne de gauche, qui indiquent les performances, montrent que les différents paramètres ont peu d'influence sur les performances du système lorsqu'elles sont mesurées par F.

Pour les trois premiers thèmes, le nombre de documents non pertinents semble avoir peu d'importance ; cependant, comme le nombre de descripteurs sélectionnés augmente avec ce paramètre, il est préférable de le limiter pour obtenir des modèles plus parcimonieux.

Pour le thème *lumber* les performances sont très variables du fait du faible nombre de documents pertinents sur la base de test : la moindre modification entraîne un grand changement dans le calcul des performances. Supposons, par exemple, que sur la base de test un filtre sélectionne six documents parmi lesquels quatre sont pertinents, la précision est alors de 66,7 % (4/6). Si l'on échange maintenant un document pertinent contre un document non pertinent, la précision devient 50 % (3/6) ce qui représente une différence de 17 points. Par conséquent les courbes présentent des fluctuations importantes, et il est difficile de tirer des conclusions claires.

La mesure du *F* caractérise un point du comportement d'un système, mais peut masquer des différences, comme nous l'avons montré dans le chapitre 4. Afin de préciser l'influence du nombre initial de descripteurs pour un nombre de documents non pertinents fixés, la Figure 7.5 présente l'évolution des courbes rappel-précision interpolée afin de préciser certains points des courbes de la Figure 7.4.
Les courbes ont été tracées avec un nombre de documents non pertinents fixé à 1000 pour le thème *interest*, à 500 pour les thèmes *oilseed*, *nat-gas*, et *sorghum* et à 100 pour le thème *lumber*. Le nombre initial de descripteurs *Q* varie d'une courbe à l'autre ; pour des raisons de lisibilité, seules les courbes correspondant à 10, 50, 200 et 400 descripteurs initiaux sont représentées.

Pour le thème *interest*, les courbes de la Figure 7.4 montrent des performances égales pour 1000 documents non pertinents en fonction du nombre initial de descripteurs. Les courbes rappel-précision précisent ce résultat et soulignent le fait que le filtre issu de la sélection des descripteurs avec 200 descripteurs initiaux est en fait supérieur aux autres, même s'ils ont tous la même valeur de *F*.
Pour le thème *oilseed* les meilleurs résultats sont obtenus avec 50 descripteurs initiaux.
Pour le thème *nat-gas*, les meilleures performances sont obtenues avec 400 descripteurs initiaux, mais les plus mauvais résultats avec 200 descripteurs initiaux. Le modèle issu de 50 descripteurs initiaux obtient des performances entre les deux.

Pour le thème *sorghum*, les meilleures performances sont obtenues avec 400 descripteurs initiaux puis avec 200. Il faut noter que, sur cette courbe, le modèle issu de 400 descripteurs initiaux apparaît comme étant supérieur notamment grâce à la précision obtenue lorsque le rappel vaut 0,8.

Les performances sur le thème *lumber* sont trop variables pour que l'on puisse en tirer des conclusions claires.

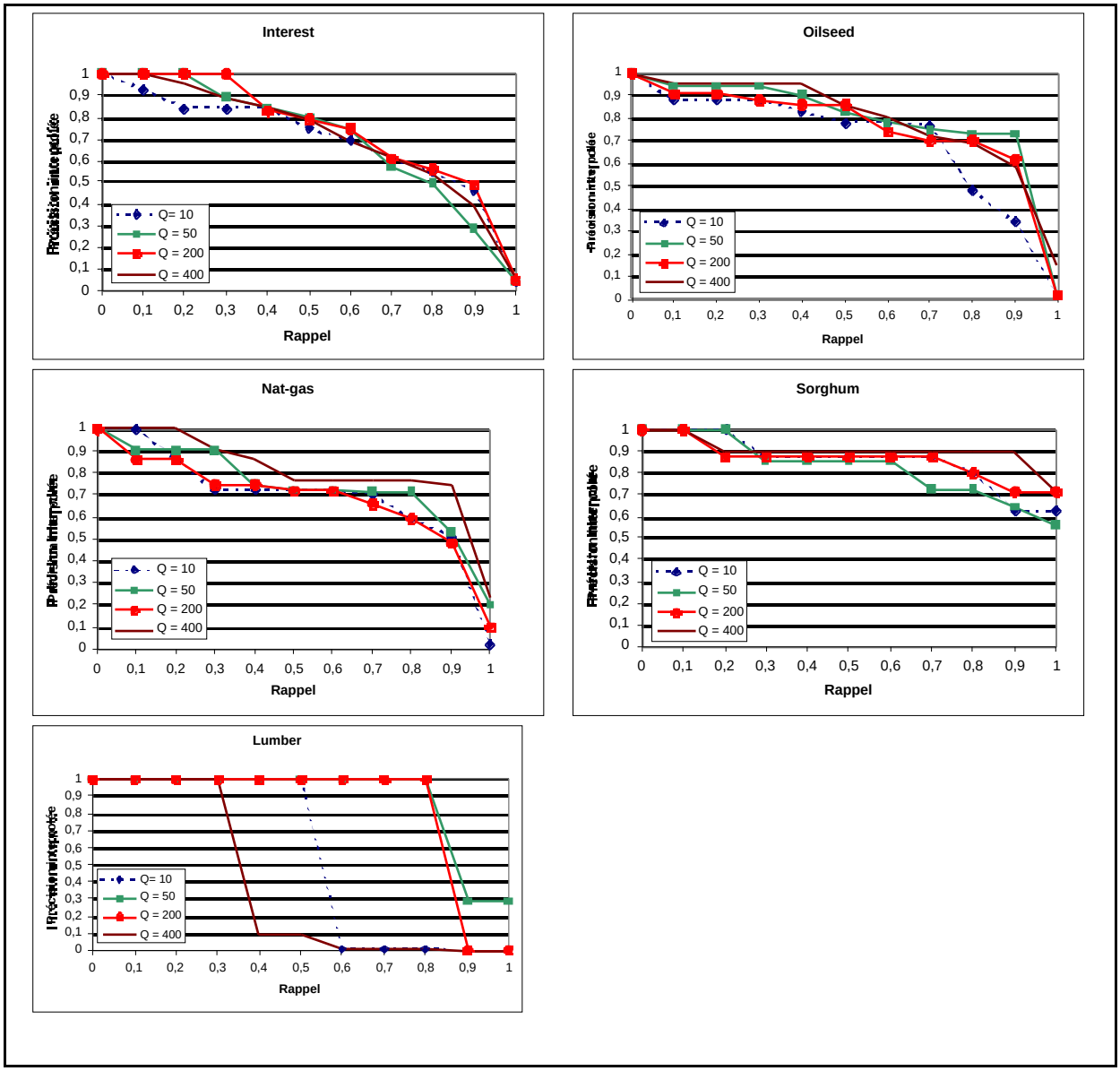

Figure 7.5 : *Courbes rappel-précision interpolée en fonction du nombre de descripteurs initiaux. Pour le thème interest le nombre de documents non pertinents est fixé à 1000, pour les thèmes oilseed, nat-gas et sorghum ce nombre est fixé à 500 et pour le thème lumber il est fixé à 100.*

L'ensemble de ces résultats montre l'existence d'une corrélation entre les deux paramètres étudiés et le nombre de descripteurs sélectionnés. En revanche, l'impact de ces paramètres sur les performances n'apparaît pas clairement et il est nécessaire d'élargir l'étude à un plus grand nombre de thèmes.

7.2.2.5 Moyenne sur les thèmes 1 à 60 du corpus Reuters

Pour mieux comprendre l'influence du nombre initial de descripteurs, nous avons réalisé une expérience sur les thèmes du corpus Reuters comprenant plus de dix documents pertinents, c'est-à-dire sur les soixante premiers thèmes.

Le modèle est toujours celui de la Figure 7.2 avec un hyperparamètre fixé à 1 pour l'apprentissage. Le nombre de documents non pertinents utilisés pour la sélection de descripteurs est de 3000 si le nombre de documents pertinents sur la base d'apprentissage est supérieur à 100, et de 500 sinon. Le nombre initial de descripteurs est de 10, 50, 100, 200 ou 400.

La Figure 7.6 présente, pour chaque valeur du nombre initial de descripteurs, le nombre moyen de descripteurs sélectionnés ainsi que les performances sur l'ensemble des thèmes. Les performances sont évaluées avec deux mesures : la moyenne de la valeur optimale de *F* obtenue sur chaque thème (macro-moyenne) et la moyenne des précisions moyennes non interpolées (UAP).

Nombre initial de descripteurs	Moyenne du nombre de descripteurs sélectionnés	UAP	*F* optimal
10	6,1	76,6	75,5
50	17,8	79,2	77,5
100	26,7	79,3	77,0
200	46,3	79,0	75,9
400	71,5	79,6	77,3

Figure 7.6 : *Macro moyenne du nombre de descripteurs sélectionnés et des performances sur les soixante premiers thèmes du corpus Reuters en fonction du nombre initial de descripteurs utilisés pour effectuer la sélection de descripteurs.*

La Figure 7.7 montre les courbes rappel-précision interpolée pour les soixante thèmes.

Ces résultats confirment que le nombre de descripteurs sélectionnés est une fonction croissante du nombre initial de descripteurs.

Avec les paramètres choisis, les moyennes varient peu en fonction du nombre Q de descripteurs initiaux ; seul le choix $Q = 10$ conduit à des résultats inférieurs sur l'ensemble des thèmes.

La solution $Q = 50$ est le meilleur choix pour l'ensemble des thèmes puisqu'elle représente le meilleur compromis entre parcimonie et performance.

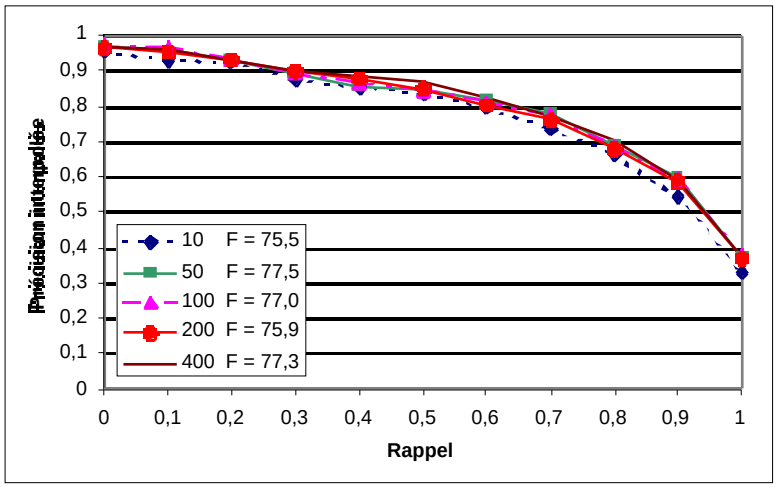

Figure 7.7 : *Courbes rappel-précision interpolée pour l'ensemble des soixante premiers thèmes du corpus Reuters en fonction du nombre initial de descripteurs utilisés pour la sélection des descripteurs. La valeur de F correspondante à chaque courbe est indiquée.*

Ces résultats sont des moyennes obtenues sur les thèmes 1 à 60, mais les performances calculées sur les sept premiers thèmes, caractérisés par un nombre de documents pertinents sur la base d'apprentissage supérieur à 300 (Figure 7.8), conduisent à des observations différentes. Dans ce cas, le meilleur compromis entre performance et parcimonie est obtenu avec 100 descripteurs initiaux. Les modèles issus de dix descripteurs initiaux sont nettement plus mauvais et contrairement à la Figure 7.6, les modèles issus de 400 descripteurs initiaux obtiennent de bonnes performances.

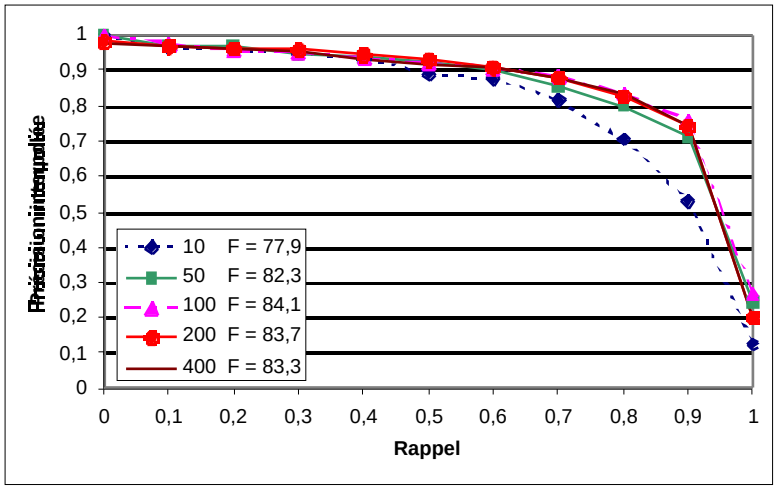

Figure 7.8 : *Courbes rappel-précision interpolée pour les sept premiers thèmes du corpus Reuters (plus de 300 documents pertinents sur la base d'apprentissage) en fonction du nombre initial de descripteurs utilisés pour la sélection des descripteurs.*

De même, les courbes de la Figure 7.9 qui présentent les résultats pour les quinze derniers thèmes (moins de vingt documents pertinents), montrent que pour obtenir les meilleures performances sur ces thèmes, il est nécessaire de limiter le nombre de descripteurs initiaux à cinquante ou dix.

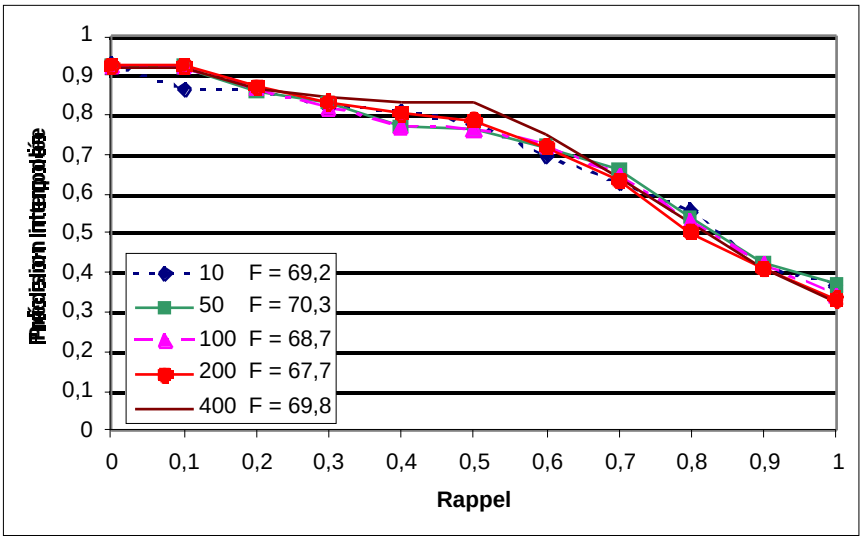

Figure 7.9 : *Courbe rappel précision interpolée pour les quinze derniers thèmes du corpus Reuters (moins de vingt documents pertinents sur la base d'apprentissage) en fonction du nombre initial de descripteurs utilisés pour effectuer la sélection de descripteurs.*

L'ensemble de ces résultats prouve que le nombre *Q* de descripteurs initiaux doit être adapté au nombre de documents pertinents. Il apparaît qu'un bon compromis consiste à choisir Q = 100

lorsque le nombre de documents pertinents est supérieur à 100, et Q = 50 lorsque le nombre de documents pertinents est inférieur à 100.

La Figure 7.10 montre les performances moyennes obtenues avec cette méthode (*Q* optimisé) par rapport à *Q* constant fixé à cinquante pour l'ensemble des thèmes.

	Moyenne du nombre de descripteurs sélectionnés	UAP	*F* optimal
Q = 50	17,8	79,2	77,5
Q optimisé	20,2	79,4	77,7

Figure 7.10 : *Comparaison entre Q = 50 et Q optimisé sur l'ensemble des soixante thèmes.*

Les différences ne sont pas très grandes entre les deux méthodes, car "l'optimisation" concerne peu de thèmes et les moyennes sur l'ensemble des thèmes masquent les différences.

La Figure 7.11 reprend la même comparaison, mais uniquement sur les dix premiers thèmes, et montre, dans ce cas, la supériorité des résultats.

	Nombre de descripteurs sélectionnés	UAP	*F* optimal
Q = 50	22,6	89,0	84,1
Q optimisé	32,2	89,7	85,3

Figure 7.11 : *Comparaison entre Q = 50 et Q optimisé sur les dix premiers thèmes.*

7.2.2.6 Influence du choix des documents non pertinents

Comme les documents non pertinents sont sélectionnés aléatoirement, il est nécessaire d'étudier la variabilité des résultats en fonction du choix de ces documents.

En conséquence, pour chacun des 5 thèmes considérés au paragraphe 7.2.2.2, 500 expériences différentes ont été effectuées, pour lesquelles le nombre de descripteur initial est systématiquement fixé à 100 et, pour chaque expérience, les documents non pertinents sont sélectionnés aléatoirement. Pour une première série d'expériences, le nombre de documents non pertinents est fixé à 1000 (Figure 7.12) et dans une deuxième série, ce nombre est fixé à 500 (Figure 7.13). Pour chacune des expériences, les valeurs moyennes du nombre de descripteurs sélectionnés, des précisions moyennes non interpolées et des mesures *F* optimales sont calculées ainsi que les écarts types.

Les résultats montrent que la dispersion des résultats augmente au fil des thèmes. L'augmentation de la dispersion s'explique car, sur le corpus Reuters, les thèmes qui ont peu de documents pertinents sur la base d'apprentissage en ont également peu sur la base de test : :pour ces thèmes, de petites modifications entraînent des variations importantes des performances.

Cependant, excepté pour le thème *lumber,* la dispersion des résultats reste faible, si bien que les résultats ne dépendent pas beaucoup du choix des documents non pertinents.

Les résultats confirment qu'il faut adapter le nombre de documents non pertinents utilisés pour la sélection de descripteurs au nombre de documents pertinents puisque, pour le thème *interest,* les meilleurs résultats sont obtenus avec 1000 documents non pertinents ; pour les

thèmes *oilseed* et *nat-gas*, les résultats sont comparables entre les deux expériences, mais les modèles sont plus parcimonieux avec 500 documents non pertinents. Pour le thème *sorghum*, les conclusions sont moins claires et les performances sont très proches avec des nombres de descripteurs sélectionnés très proches également. Enfin, pour le thème *lumber*, les écarts types sont trop élevés pour tirer des conclusions.

		Nombre de descripteurs sélectionnés	UAP	*F* optimal
Interest	Moyenne	19,4	73,0	67,5
	Ecart-type	2,0	1,2	1,3
Oilseed	Moyenne	24,5	77,6	78,4
	Ecart-type	2,6	1,3	1,5
Nat-gas	Moyenne	19,0	68,5	67,2
	Ecart-type	2,8	2,6	2,4
Sorghum	Moyenne	23,6	80,2	79,7
	Ecart-type	4,7	2,6	3,6
Lumber	Moyenne	19,0	58,0	53,2
	Ecart-type	5,8	9,8	6,5

Figure 7.12 : *Nombre de descripteurs moyens et performances moyennes pour 500 tirages différents de 1000 documents non pertinents.*

		Nombre de descripteurs sélectionnés	UAP	*F* optimal
Interest	Moyenne	15,2	72,6	66,3
	Ecart-type	1,6	1,4	1,4
Oilseed	Moyenne	20,2	77,3	78,8
	Ecart-type	2,3	1,3	1,3
Nat-gas	Moyenne	16,0	68,7	68,9
	Ecart-type	2,9	3,2	2,9
Sorghum	Moyenne	24,1	79,5	79,5
	Ecart-type	5,2	2,5	3,5
Lumber	Moyenne	22,6	55,1	52,7
	Ecart-type	7,0	13,1	6,8

Figure 7.13 : *Nombre de descripteurs moyens et performances moyennes pour 500 tirages différents de 500 documents non pertinents.*

7.2.2.7 Cas des catégories avec très peu de documents

Dans le cas du corpus Reuters, les trente derniers thèmes possèdent moins de dix documents pertinents, et les vingt-cinq derniers ont un nombre de documents pertinents inférieur ou égal à cinq. Dans ce cas, il est probable que la plupart des mots issus du vocabulaire spécifique apparaissent systématiquement ensemble dans les documents pertinents et sont donc très corrélés ; comme la méthode d'orthogonalisation de Gram-Schmidt exploite ces corrélations, elle peut être conduite à sélectionner très peu de descripteurs. De plus, les résultats du paragraphe précédent ont montré que, pour les thèmes possédant très peu de documents pertinents, les résultats dépendaient beaucoup des documents non pertinents utilisés.

Par conséquent, nous avons comparé les résultats obtenus par la méthode d'orthogonalisation de Gram-Schmidt et ceux obtenus en considérant uniquement les dix premiers mots trouvés par la méthode du vocabulaire spécifique sur les trente derniers thèmes du corpus Reuters. Pour la méthode de Gram-Schmidt, Q est fixé à dix, et 500 documents non pertinents sont utilisés.

Les résultats sont présentés à la Figure 7.14 pour les trente derniers thèmes (ensemble des thèmes avec moins de dix documents pertinents) et les vingt derniers thèmes (ensemble des thèmes avec moins de cinq documents pertinents) et sont calculés grâce à la moyenne des mesures de F pour chaque thème et par les moyennes des précisions moyennes non interpolées (UAP).

	30 derniers thèmes			20 derniers thèmes		
	Nombre de descripteurs sélectionnés	UAP	*F*	Nombre de descripteurs sélectionnés	UAP	*F*
Vocabulaire spécifique	10	59,0	43,8	10	59,0	36,7
Gram-Schmidt	14,9	50,2	44,0	12,1	38,9	33,3

Figure 7.14 : *Comparaison des méthodes de sélection de descripteurs sur les thèmes possédant peu de documents pertinents.*

La méthode du vocabulaire spécifique s'avère meilleure pour l'ensemble des thèmes possédant peu de documents pertinents, d'autant plus qu'elle implique des modèles plus parcimonieux et qu'elle nécessite moins de calculs.

Cependant, contrairement aux autres résultats obtenus, les conclusions sont légèrement différentes selon la mesure : avec la précision moyenne non interpolée, les différences entre les deux approches sont très élevées, alors que la mesure de *F* met en avant des écarts plus faibles. Cette différence s'explique par certains thèmes comme le thème *dfl* (le thème 82 de notre liste) dont les résultats sont présentés à la Figure 7.15, et qui ne possède qu'un seul document pertinent sur la base de test.

	UAP	*F*
Vocabulaire spécifique	100	0,06
Gram-Schmidt	0,06	0,06

Figure 7.15 : *Résultats pour le thème dfl.*

Pour ce thème, les mesures de *F* ont la même valeur (la valeur minimale) pour les deux approches, mais la précision moyenne non interpolée est très différente d'une méthode à l'autre, puisqu'elle est maximale avec la méthode du vocabulaire spécifique et minimale avec la méthode de Gram-Schmidt. Pour comprendre cette différence, la Figure 7.16 montre les cinq probabilités les plus élevées de la base de test lorsque le vecteur d'entrée du modèle est sélectionné par la méthode du vocabulaire spécifique : si le document pertinent est bien classé en première position, sa probabilité est trop faible pour qu'il soit sélectionné par une méthode de seuil.

Dans ce cas, la mesure de la précision moyenne non interpolée n'est plus une bonne mesure, car, malgré la valeur maximale obtenue, le système ne peut pas exploiter le bon classement.

Sortie du réseau	Pertinent
0,080	oui
0,060	non
0,059	non
0,057	non
0,052	non

Figure 7.16 : *Cinq premières probabilités pour le thème dfl.*

7.2.3 Conclusion sur la mise en œuvre de Gram-Schmidt

L'ensemble des expériences effectuées sur la sélection de descripteurs a permis de tirer plusieurs enseignements, et notamment que le nombre de documents non pertinents ainsi que le nombre initial de descripteurs doivent être choisis en fonction du nombre de documents pertinents sur la base d'apprentissage.

Plus précisément :

- Le nombre de descripteurs sélectionnés est une fonction croissante du nombre de documents non pertinents et du nombre initial de descripteurs. Le facteur prédominant est le nombre initial de descripteurs.
- Le nombre de documents non pertinents et le nombre initial de descripteurs doivent être limités en fonction du nombre de documents pertinents sur la base d'apprentissage, afin d'obtenir le meilleur compromis entre performance et parcimonie.

Après la sélection de descripteurs, il reste à déterminer les caractéristiques de la base d'apprentissage d'une part, et le nombre de neurones cachés de l'architecture neuronale d'autre part.

7.3 Choix des documents non pertinents pour la base d'apprentissage

Comme pour l'étape de sélection de descripteurs, il est nécessaire de choisir des documents pertinents et des documents non pertinents pour constituer une base d'apprentissage.
Comme précédemment l'ensemble des documents pertinents disponibles est retenu et il reste à choisir un sous-ensemble de documents non pertinents, ce sous-ensemble n'étant pas nécessairement le même que celui utilisé pour la sélection des descripteurs.
L'expérience suivante étudie l'influence de la constitution de la base d'apprentissage sur les performances et plus précisément l'impact du nombre de documents non pertinents. Les cinq thèmes du paragraphe précédent sont étudiés ; la méthode de sélection de descripteurs tient compte des résultats obtenus précédemment : elle est réalisée avec les paramètres de la Figure 7.17.

	Nombre de documents non pertinents pour la sélection de descripteurs.	Nombre de descripteurs initiaux	Nombre de descripteurs sélectionnés
interest	3000	200	36
oilseed	500	50	18
nat-gas	500	50	12
sorghum	500	50	19
lumber	100	50	17

Figure 7.17 : *Paramètres de la sélection de descripteurs.*

Comme les descripteurs sélectionnés pour représenter les textes sont en nombre réduit, la plupart des textes ne comportent aucun de ces mots. En d'autres termes, avec la représentation choisie, la majorité des textes sont tout simplement des vecteurs nuls. Par construction, les documents pertinents ne doivent pas être des vecteurs nuls puisque les descripteurs ont été choisis dans l'ensemble représentatif de ce sous-ensemble.

La Figure 7.18 donne, pour les 5 thèmes étudiés, la proportion de vecteurs non nuls parmi 5000 documents non pertinents sélectionnés aléatoirement, représentés par les descripteurs sélectionnés avec les paramètres de la Figure 7.17.

	Nombre de documents non pertinents différents du vecteur nul.	Proportion de vecteur non nuls
interest	1924	38,4%
oilseed	664	13,2%
nat-gas	581	11,6%
sorghum	783	15,6%
lumber	249	5,0%

Figure 7.18 : *Nombre et proportion de vecteurs non nuls sur 5000 documents non pertinents sélectionnés aléatoirement avec les descripteurs de la Figure 7.17.*

Les résultats montrent que, quel que soit le thème, une majorité de documents non pertinents sont en fait des vecteurs nuls, ce qui signifie, par exemple, que pour le thème *lumber*, 95 % des documents non pertinents apportent exactement la même information : un document qui ne contient aucun des descripteurs n'est pas pertinent.

Une expérience a été effectuée pour chaque thème avec le modèle de la Figure 7.2 et un hyperparamètre fixé à 1, en faisant varier le nombre de documents non pertinents dans la base d'apprentissage : 1000, 2000, ou un ensemble de documents non pertinents tel qu'aucun ne soit représenté par un vecteur nul.

Les courbes rappel-précision interpolée sont présentées à la Figure 7.19, pour un thème donné ; chaque courbe correspond à un choix différent de l'ensemble des documents non pertinents constituant la base d'apprentissage.

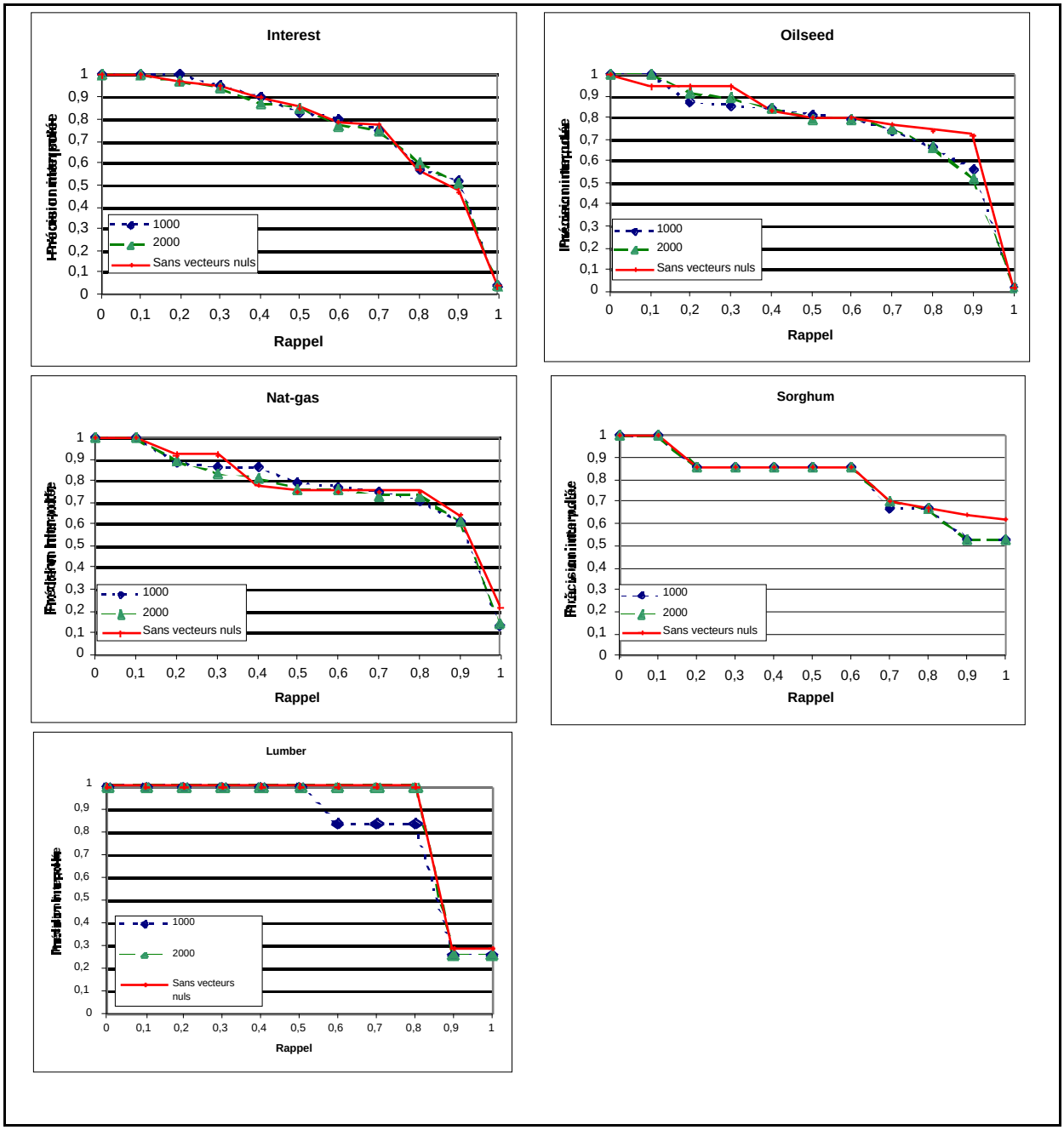

Figure 7.19 : *Courbes rappel-précision interpolée en fonction du choix des dépêches non pertinentes dans la base d'apprentissage.*

Une expérience similaire a été menée sur les cinquante thèmes du corpus TREC-8 où, pour chaque thème, deux bases d'apprentissage différentes sont testées. Dans les deux cas, tous les documents pertinents disponibles sont pris en considération, mais dans le premier cas, le sous-ensemble des documents non pertinents est constitué de 1300 documents sélectionnés aléatoirement et dans le deuxième cas, le sous-ensemble des textes non pertinents est constitué de vecteurs non nuls compte tenu des descripteurs sélectionnés.

La Figure 7.20 est la courbe rappel-précision interpolée obtenue pour les cinquante thèmes dans chacun des cas. Cette courbe est obtenue en calculant pour chaque valeur de rappel r, la moyenne des précisions interpolées obtenues pour chaque thème pour cette même valeur r.

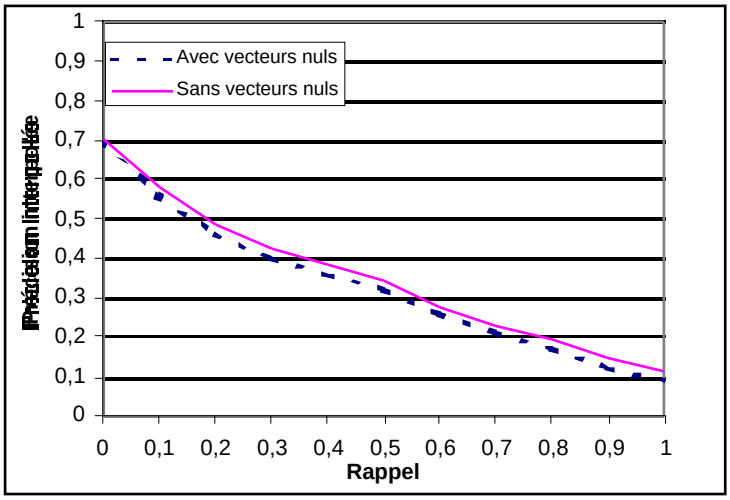

Figure 7.20 : *Courbes rappel-précision interpolée obtenues pour l'ensemble des cinquante thèmes du corpus TREC-8. Chaque point de la courbe est obtenu en effectuant la moyenne de chacun des thèmes.*

Ces expériences, menées sur des corpus différents, montrent qu'il est préférable d'utiliser des bases d'apprentissage qui ne comportent pas de vecteur nul dans le sous-ensemble des documents non pertinents.

Cette approche présente l'avantage de créer des bases d'apprentissage de taille réduite, donc de limiter les temps de calculs. De plus, elle permet de s'affranchir du choix de la taille de la base en proposant un choix systématique adapté à chaque thème.

7.4 Étude du nombre de neurones cachés

Jusqu'ici, toutes les expériences ont été effectuées avec aucun neurone caché, c'est-à-dire avec une régression logistique. Un tel modèle est appelé modèle linéaire car la surface de séparation

obtenue est un hyperplan dans l'espace des descripteurs. En ajoutant des neurones cachés dans la couche cachée, on crée des surfaces de séparation plus complexes qui peuvent améliorer les performances de classification.

7.4.1 Théorème de Cover

Avant d'essayer d'ajouter des neurones cachés à la structure neuronale, il faut s'assurer que le problème de classification traité n'est pas intrinsèquement linéairement séparable, auquel cas il est inutile d'essayer d'ajouter des non-linéarités à la surface de séparation.
Le théorème de Cover [Cover, 1965] précise les conditions dans lesquelles un problème de classification est toujours linéairement séparable.

Son énoncé est le suivant :
Soit un problème de classification à deux classes, dont les deux classes sont équiprobables, comprenant N exemples dans un espace de dimension d, alors :

- Si $N \leq d + 1$, n'importe quelle dichotomie des exemples est linéairement séparable.
- Si $N = 2(d + 1)$, n'importe quelle dichotomie est linéairement séparable avec une probabilité de 0,5.

Pour les problèmes de classification traités, le nombre d'exemples pertinents peut être inférieur au nombre de descripteurs utilisés dans la représentation des textes. Cependant, le nombre de documents non pertinents disponibles sur les corpus étudiés est très élevé, comme on l'a précisé au paragraphe 7.2.2.1. Même si une grande partie des documents non pertinents est représentée par le vecteur nul, il est toujours possible de choisir un nombre de documents non pertinents suffisamment élevé pour être dans les conditions du théorème de Cover où le problème de classification n'est pas, *a priori*, linéairement séparable.
Il n'est donc pas évident, *a priori*, que le meilleur classifieur soit un séparateur linéaire.

7.4.2 Variations des performances en fonction du nombre de neurones cachés

Pour faire des comparaisons équitables, il est important de mener correctement l'apprentissage des réseaux de neurones contenant des neurones cachés : il faut utiliser une méthode de *weight decay* et en déterminer les hyperparamètres. Or comme il a été précisé dans le chapitre 6, pour

les réseaux de neurones à couches, il est préférable d'utiliser plusieurs hyperparamètres dont le réglage peut être délicat.

Cependant, comme on le verra dans le chapitre 9, il est possible de fabriquer des bases d'apprentissage de grande taille sur le corpus AFP qui ne nécessitent donc pas nécessairement de méthodes de régularisation. Les différentes expériences qui ont pu être faites sur ces thèmes du corpus de l'AFP ont montré que l'ajout de neurones cachés n'améliorait pas les résultats [Stricker *et al*., 2000].

Sur l'ensemble des cinquante thèmes du corpus TREC-8, des architectures à 0, 1 ou 2 neurones cachés ont été testées avec une méthode d'arrêt prématuré explicitée au paragraphe 7.6.2.
Pour ces différentes architectures, les vecteurs d'entrées des réseaux de neurones sont identiques et les résultats de la Figure 7.21 sont calculés sur la base de test en moyennant les précisions moyennes non interpolées de chaque thème.

	0 neurone caché	1 neurone caché	2 neurones cachés
Moyenne des précisions moyennes non interpolées	33,18	33,09	28,92

Figure 7.21 : *Comparaison des résultats obtenus sur le corpus TREC-8 avec différentes architectures. Les thèmes sans documents pertinents sur la base de test ne sont pas pris en considération.*

Les résultats montrent que, en moyenne, le modèle linéaire donne de meilleurs résultats que le modèle non linéaire.
Il existe cependant des différences thème par thème : sur la Figure 7.22, chaque thème est représenté par un point, dont l'abscisse est sa précision moyenne non interpolée calculée avec le modèle linéaire, et dont l'ordonnée est celle issue du modèle avec deux neurones cachés. La droite $y = x$ sépare le plan en deux : si un point est dans le demi-plan inférieur, le modèle linéaire est meilleur et dans le cas contraire, le modèle avec deux neurones cachés est meilleur.

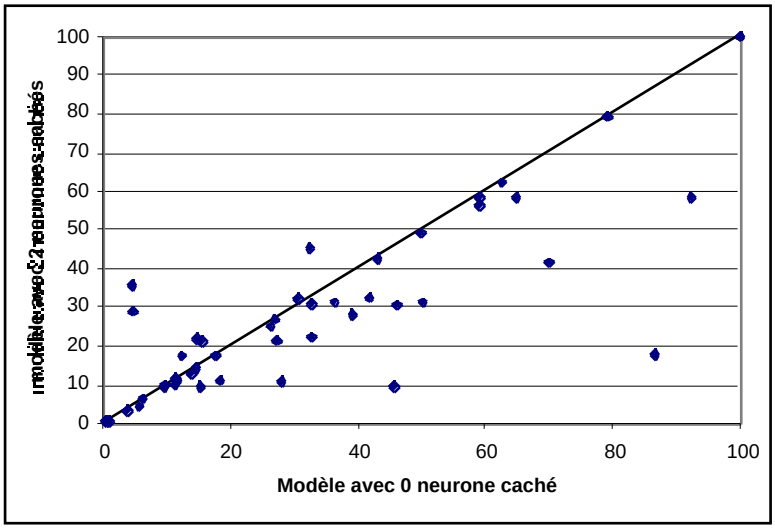

Figure 7.22 : *Résultat thème par thème pour les cinquante thèmes du corpus TREC-8. Chaque point représente un thème, son abscisse est l'interpolation moyenne non interpolée obtenue avec aucun neurone caché et son ordonnée est celle obtenue avec deux neurones cachés.*

Les résultats montrent que, pour la grande majorité des thèmes le modèle avec aucun neurone caché est meilleur que le modèle avec deux neurones cachés. Les trois thèmes pour lesquels le modèle à 2 neurones cachés apporte des améliorations significatives sont les thèmes 362, 382, et 393 qui ont respectivement 4, 5, et 5 documents pertinents sur la base de test et pour lesquels la mesure est donc très sensible.

Finalement, ces résultats montrent que le fait d'ajouter des neurones cachés n'améliore pas les performances en moyenne, et il semble même vain de chercher à optimiser ce nombre pour chaque thème puisque, dans la grande majorité des cas, le modèle le plus simple est le meilleur.

7.4.3 Conclusion sur le nombre optimal de neurones cachés

L'ensemble de ces résultats, obtenus sur des corpus différents, montre que l'ajout de neurones cachés et donc l'utilisation de surfaces de séparation non linéaires n'améliore pas les résultats, et, la plupart du temps, les diminue.

Cette observation est assez étonnante d'autant plus que le langage naturel est *a priori* complexe. Cependant, ces observations confirment les résultats obtenus par d'autres auteurs qui utilisent des réseaux de neurones pour faire de la catégorisation de textes. Ces travaux sont rapidement décrits ci-dessous.

[Wiener, 1993] utilise des réseaux de neurones avec la même architecture que celle présentée à la Figure 7.1, sur l'ancienne version du corpus Reuters (Reuters-22173). Dans cette étude, plusieurs architectures avec des nombres de neurones cachés différents sont testées, allant jusqu'à 6 neurones cachés. Pour ces différentes études, les sélections de descripteurs ont été effectuées soit par des méthodes de sélections de termes comme la méthode du chi-2, soit par des variantes de la méthode *latent semantic indexing* [Deerwester *et al.*, 1990] qui proposent une approche différente des méthodes de sélections précédentes, s'appuyant sur une décomposition en valeurs singulières. Les apprentissages sont effectués en conjuguant une méthode d'arrêt prématuré avec une méthode de pénalisation des poids

D'après les résultats obtenus dans cette étude, les réseaux de neurones avec neurones cachés n'obtiennent pas de résultats significativement meilleurs que les réseaux sans neurone caché et ce, quelle que soit la méthode de sélection de descripteurs :

"*There was surprisingly little gain in effectiveness of the non linear networks over the linear networks"*

Afin de prouver que ses algorithmes ne sont pas en cause, il invente des thèmes artificiels qui sont des compositions de thèmes existants, et qui nécessitent des modèles non linéaires. Dans ce cas, les réseaux de neurones avec des neurones cachés obtiennent bien de meilleurs résultats que les réseaux linéaires, ce qui tend à prouver qu'il ne s'agit pas d'un problème algorithmique, mais bien d'un problème structurel.

[Schütze *et al.*, 1995] ont également étudié une approche neuronale pour effectuer de la catégorisation de textes sur les corpus issus de TREC-2 et TREC-3. Dans leurs expériences, la sélection de descripteurs est effectuée comme précédemment, soit par la méthode *latent semantic indexing,* soit par la méthode du chi-2. Ils comparent également une architecture ne contenant pas de neurone caché avec une architecture contenant trois neurones cachés, l'apprentissage s'effectuant avec la méthode de l'arrêt prématuré. Pour cette expérience également, l'architecture comprenant des neurones cachés n'apporte pas d'améliorations significatives par rapport à l'architecture sans neurone caché :

"*It is safe to conclude that the non-linear components to the neural network provides absolutely no advantage*"

[Yang et Liu, 1999] utilisent des réseaux de neurones avec des neurones cachés sur le corpus Reuters-21578, et fixent le nombre de neurones cachés grâce à un ensemble de validation (dont ils ne précisent pas la composition). Ils testent différentes architectures avec 16, 64 ou 160 neurones cachés et grâce à la base de validation, ils choisissent un nombre de neurones cachés égal à 64. Pour leurs expériences, ils sélectionnent 1000 descripteurs grâce au calcul de l'information mutuelle, et par conséquent leurs réseaux de neurones comportent environ 64.000 poids à déterminer.

Ils ne font pas ici de comparaison avec un simple modèle linéaire, mais il faut noter que les résultats obtenus avec leurs réseaux sont assez faibles en comparaison des autres méthodes, si bien qu'il est impossible de dire si les neurones cachés ont permis une amélioration des résultats.

7.4.4 Vers une représentation plus élaborée

Pourquoi les réseaux linéaires obtiennent-ils de meilleurs résultats que les réseaux non linéaires ?

Une des explications provient des méthodes de sélection des descripteurs qui sélectionnent les descripteurs un par un sans tenir compte de leurs interactions éventuelles en particulier des méthodes comme l'information mutuelle ou le chi-2, cette critique étant moins vraie pour la méthode d'orthogonalisation de Gram-Schmidt et la méthode *latent semantic indexing*.

Cependant toutes les études citées, ainsi que la nôtre, conduisent aux mêmes conclusions, bien qu'elles utilisent des corpus différents qui regroupent une grande variété de situations.

Il semble que la représentation des textes par le modèle vectoriel conduise à des ensembles d'exemples qui admettent comme meilleur classifieur un séparateur linéaire. Bien entendu, cette situation est une conséquence directe de la représentation choisie, et il n'est pas certain qu'elle se retrouve pour des représentations plus élaborées telles que celles que nous décrirons dans le chapitre 8.

7.5 Mise en œuvre sur l'ensemble du corpus Reuters

Ce paragraphe présente les résultats obtenus sur l'ensemble des 90 catégories du corpus Reuters à partir de l'ensemble des remarques effectuées dans ce chapitre. Ces résultats peuvent être comparés à d'autres résultats de la littérature.

7.5.1 Choix des paramètres du modèle

Pour la sélection des descripteurs, les paramètres choisis sont les suivants :

- Si le nombre de documents pertinents est supérieur à cent, on choisit 3000 documents non pertinents et 100 descripteurs initiaux.
- Si le nombre de documents pertinents est supérieur à dix, on choisit 500 documents non pertinents et 50 descripteurs initiaux.
- Lorsque le nombre de descripteurs pertinents est inférieur à dix, les dix premiers descripteurs trouvés par la méthode du vocabulaire spécifique sont sélectionnés.

Le modèle est une régression logistique, et l'apprentissage est effectué avec la méthode du *weight decay* et un hyperparamètre fixé à 1 ; la base d'apprentissage est constituée de tous les documents pertinents, et de documents non pertinents pour lesquels la représentation est différente du vecteur nul.

7.5.2 Performances sur l'ensemble du corpus Reuters

Plusieurs mesures ont été calculées pour faciliter les comparaisons avec les autres travaux : la précision moyenne non interpolée (UAP), la moyenne sur 11 points (11-pt), la mesure de F (selon les définitions exposées au chapitre 4).

La Figure 7.23 présente les macro-moyennes pour chacune de ces mesures sur l'ensemble des thèmes. Les moyennes sont aussi calculées pour plusieurs sous-ensembles de thèmes groupés en fonction du nombre de documents pertinents sur la base d'apprentissage :

- Les dix premiers thèmes, car beaucoup d'auteurs publient des résultats sur ces thèmes qui ont le plus de documents pertinents.
- Les thèmes 11 à 40, pour les thèmes possédant plus de vingt-cinq documents pertinents.

- Les thèmes 41 à 60, pour les thèmes ayant entre dix et vingt-cinq documents pertinents.
- Les thèmes 61 à 90, pour les thèmes possédant moins de dix documents pertinents, et pour lesquels la sélection de descripteurs est effectuée différemment des autres.

	Nombre de descripteurs	UAP	11-pt	*F*
Moyenne sur l'ensemble des thèmes	16,8	72,6	72,8	66,4
Moyenne pour les thèmes 1 à 10	32,2	89,7	87,2	85,3
Moyenne pour les thèmes 11 à 40	18,3	83,2	83,6	81,4
Moyenne pour les thèmes 41 à 60	17,4	69,2	69,1	69,1
Moyenne pour les thèmes 61 à 90	10,0	59,0	59,6	43,8

Figure 7.23 : *Ensemble des résultats sur le corpus Reuters.*

Ces expériences montrent, que, en moyenne, les performances décroissent lorsque nombre de documents pertinents de la base d'apprentissage diminue.

7.5.3 Influence de la valeur de l'hyperparamètre

Pour tester la sensibilité des résultats à la valeur de l'hyperparamètre, plusieurs valeurs ont été testées. La variation de la précision moyenne non interpolée en fonction de la valeur de l'hyperparamètre α est présentée à la Figure 7.24. Ces résultats montrent que la valeur choisie a peu d'influence sur les résultats tant qu'elle n'est pas trop élevée. La valeur 5 donne clairement de moins bons résultats sur les catégories comprenant peu d'exemples d'apprentissage, car dans ce cas, le terme de pénalisation prend trop d'importance par rapport au terme d'entropie croisée, et les poids tendent vers zéro.

	$\alpha = 0,1$	$\alpha = 0,5$	$\alpha = 1,0$	$\alpha = 5,0$
Moyenne sur l'ensemble des thèmes	72,4	72,5	72,6	68,3

Moyenne pour les thèmes 1 à 10	89,2	89,5	89,7	89,6
Moyenne pour les thèmes 11 à 40	82,0	82,9	83,2	82,3
Moyenne pour les thèmes 41 à 60	70,2	69,4	69,2	65,9
Moyenne pour les thèmes 61 à 90	59,1	59,0	59,0	49,3

Figure 7.24 : *Comparaison des performances mesurées avec la précision moyenne non interpolée en fonction de la valeur de l'hyperparamètre.*

7.5.4 Hyperparamètre déterminé par la méthode d'intégration

Jusqu'ici l'hyperparamètre était fixé à une valeur constante pendant l'apprentissage, mais l'approche bayesienne présentée au chapitre 6 a montré que la valeur de l'hyperparamètre pouvait être fixée pendant l'apprentissage.

La méthode d'intégration décrite au chapitre 6 a été utilisée : si p est le nombre de poids du réseau, alors l'hyperparamètre est calculé régulièrement pendant l'apprentissage selon la formule :

$$\alpha = \frac{p}{\sum_{i=1}^{p} w_i^2}$$

La valeur initiale de l'hyperparamètre est fixée à 1.

La comparaison des performances obtenues avec cette méthode et avec un hyperparamètre fixé à 1 est présentée à la Figure 7.25.

Dans ce cas, la méthode d'intégration issue de l'approche bayesienne n'apporte pas d'amélioration : les résultats sont très proches. La méthode d'intégration nécessite cependant plus de calculs : il semble préférable de se contenter d'une valeur constante de l'hyperparamètre, d'autant plus que les résultats de la Figure 7.24 montrent que le choix de cette valeur n'est pas critique.

	α variable	α = 1
Moyenne sur l'ensemble des thèmes	71,6	72,6

Moyenne pour les thèmes 1 à 10	89,4	89,7
Moyenne pour les thèmes 11 à 40	82,2	83,2
Moyenne pour les thèmes 41 à 60	67,8	69,2
Moyenne pour les thèmes 61 à 90	58,1	59,0

Figure 7.25 : *Comparaison des performances mesurées avec la précision moyenne non interpolée en fonction de la valeur de l'hyperparamètre.*

La méthode de maximisation n'a pas été testée sur ce problème, car l'ensemble des résultats semble indiquer que la valeur de l'hyperparamètre, avec ce modèle, a peu d'influence.

7.5.5 Utilisation de racines lexicales

Jusqu'ici, les descripteurs utilisés étaient les mots tel qu'ils apparaissaient dans les textes ; chaque flexion d'un mot était considérée comme un descripteur différent.

Dans l'expérience décrite ci-dessous, tous les mots des textes sont remplacés par leur racine lexicale selon la méthode décrite au chapitre 5**,** et les fréquences d'apparition de chaque racine sur l'ensemble du corpus sont calculées**.** La détermination du vocabulaire spécifique, ainsi que la sélection de descripteurs, sont ensuite effectuées exactement comme précédemment pour chaque thème.

La Figure 7.26 montre les dix premiers descripteurs sélectionnés pour le thème *interest* lorsque les textes sont conservés tel quel ou lorsque les mots sont substitués par leur racine. La première liste contient les mots *rates* et *rate* qui deviennent la racine *rate* dans la deuxième liste. Il faut noter cependant que la liste des descripteurs sélectionnés à partir des textes utilisant les racines n'est pas identique à la liste des racines des descripteurs sélectionnés à partir des textes originaux : par exemple, le mot *opened* est sélectionné dans les descripteurs obtenus à partir des textes originaux, mais sa racine *open* ne fait pas partie des descripteurs sélectionnés à partir des textes utilisant les racines : *opened* est une forme relativement rare qui peut être discriminante, mais sa racine *open* est un mot très courant qui n'est plus discriminant.

Normal	Racine
Rate	rate
money	monei
customer	fed
prime	prime
england	england
rates	custom
band	band
bundesbank	bundesbank
discount	discount
opened	repurchas

Figure 7.26 : *Thème interest : liste des dix premiers descripteurs sélectionnés à partir des textes originaux ou à partir des textes avec les racines.*

Les performances sur l'ensemble des catégories du corpus sont présentées à la Figure 7.27 ; ces résultats sont à comparer avec les résultats obtenus à la Figure 7.23 sans l'utilisation de racines.

Quel que soit l'ensemble de catégories considérées, les résultats sont systématiquement inférieurs à ceux obtenus en conservant les mots inchangés.

	Nombre de descripteurs	UAP	11-pt	*F*
Moyenne sur l'ensemble des thèmes	14,5	71,4	71,5	66,3
Moyenne pour les thèmes 1 à 10	28,8	89,4	86,3	84,9
Moyenne pour les thèmes 11 à 40	15,1	82,5	82,7	81,0
Moyenne pour les thèmes 41 à 60	13,9	67,5	68,3	65,4
Moyenne pour les thèmes 61 à 90	10,0	57,4	58,0	46,4

Figure 7.27: *Ensemble des résultats sur le corpus Reuters avec les racines.*

Les résultats précédents sont des moyennes. La Figure 7.28 montre, pour les soixante premiers thèmes, les différences thème à thème : chaque point représente un thème, l'abscisse d'un point est la précision moyenne non interpolée obtenue sans l'utilisation de racines et son ordonnée est celle obtenue en substituant les mots par leur racine.

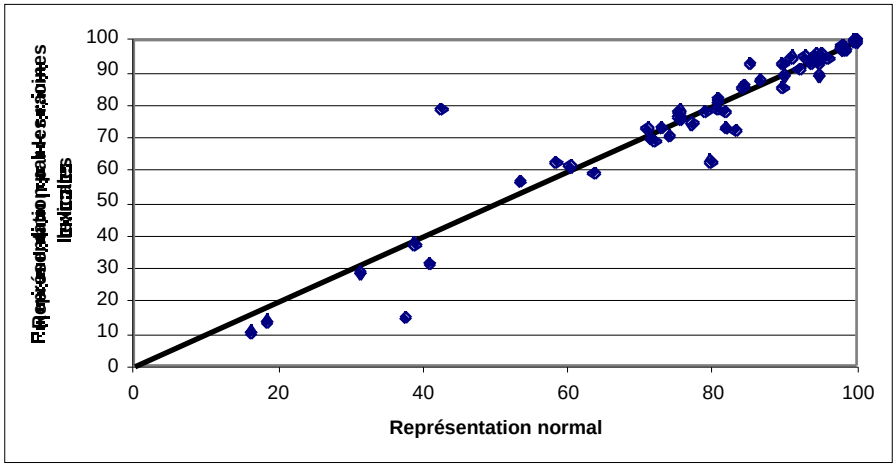

Figure 7.28 : *Comparaison des précisions moyennes non interpolées pour les soixante premiers thèmes : représentation normal ou représentation avec des racines lexicales.*

Le seul thème pour lequel il existe une grande différence est le thème *lumber* (thème 60 selon notre classement) dont la précision moyenne non interpolée passe de 0,42 à 0,78 grâce à

l'utilisation de racines. Cette différence est due à la présence du mot *wood* dans la liste des descripteurs sélectionnés lorsque l'on utilise les racines : la présence de ce mot permet de classer le document pertinent 20790 en tête de classement et, comme il n'existe que 6 documents pertinents pour ce thème, cela produit une grande différence de performance.

Pour l'ensemble des autres thèmes, les différences sont peu importantes et, selon les thèmes, certaines performances sont améliorées et d'autres dégradées. Cependant, comme l'ont montré les performances de la Figure 7.27, il semble préférable, en moyenne, de conserver les mots originaux plutôt que de les remplacer par leur racine lexicale.

L'algorithme utilisé n'est pas parfait, et certains rapprochements de mots sont injustifiés. Dans la liste des racines de la Figure 7.26 se trouve, par exemple, la racine *custom*. Or cette racine correspond aux mots *customers* (clients) et *customer* (client), mais elle correspond également au mot *customs* (douane). Donc avec l'utilisation de racines, les mots *douanes* et *clients* sont considérés comme identiques, alors que leur signification est évidemment différente. Par conséquent, selon les catégories et les descripteurs utilisés, l'utilisation de racines est susceptible soit d'améliorer, soit de dégrader, les performances de classification.

L'impact de l'utilisation des racines a beaucoup été étudié dans la communauté de la recherche d'informations ; différents algorithmes ont été utilisés, comme celui développé par [Lovins, 1968] ou d'autres fondés sur des analyses morphologiques [Hull, 1996]. Les conclusions de ces études sont parfois différentes, mais il semble que, globalement, les conclusions soient proches de nos résultats.

Dans son étude, [Harman, 1991] a testé plusieurs algorithmes, et conclut que l'utilisation de racines n'améliore pas les résultats, mais [Krovetz, 1993] a observé des améliorations significatives grâce à leur utilisation. [Hull, 1996] a également proposé une étude intensive sur l'utilisation des racines et a montré que les améliorations de performances étaient faibles en moyenne. Dans cette étude, il étudie précisément certaines requêtes et exhibe plusieurs séries de mots regroupés sous la même racine tout en ayant des sens différents comme le mot *server* de l'expression *client-server* qui devient *serve* c'est-à-dire un verbe extrêmement commun.

Dans toutes ces études, certains thèmes voient leur performance s'améliorer, d'autres voient leur performance se dégrader, si bien que, en moyenne, les améliorations éventuelles sont faibles.

Le problème de l'utilisation des racines provient essentiellement de rapprochement de mots avec des sens différents sous la même racine. Ces rapprochements fabriquent de faux synonymes et, finalement, avec notre modèle, il semble préférable de ne pas utiliser de racines.

7.5.6 Utilisation de lemmes

Le défaut principal des racines étant de regrouper trop de mots différents sous une même racine, l'utilisation de lemmes pourrait résoudre ce problème, car les rapprochements de mots sont effectués sur la base d'une analyse grammaticale grâce à l'algorithme présenté au chapitre 5.

La Figure 7.29 montre les dix premiers descripteurs sélectionnés pour le thème *interest* selon l'utilisation ou non de lemmes (la liste obtenue avec les racines est également présente pour faciliter les comparaisons). Par rapport à la liste de la Figure 7.26, *customer* n'est plus maintenant assimilé à *customs*.

Normal	**Lemme**	**Racine**
rate	rate	rate
money	customer	monei
customer	england	fed
prime	prime	prime
england	money	england
rates	band	custom
band	repurchase	band
bundesbank	discount	bundesbank
discount	feed	discount
opened	cuts	repurchas

Figure 7.29 : *Thème interest : liste des dix premiers descripteurs sélectionnés à partir des textes originaux, des lemmes et des racines lexicales.*

Les performances sur l'ensemble des catégories du corpus sont présentées à la Figure 7.30 ; ces résultats sont à comparer avec les résultats obtenus à la Figure 7.23 sans l'utilisation de lemmes, et avec les résultats de la Figure 7.27 obtenus avec les racines.

	Nombre de descripteurs	UAP	11-pt	*F*
Moyenne sur l'ensemble des thèmes	15,3	70,6	70,7	66,0
Moyenne pour les thèmes 1 à 10	30,4	89,5	86,8	84,7
Moyenne pour les thèmes 11 à 40	15,6	82,8	82,8	80,6
Moyenne pour les thèmes 41 à 60	15,5	66,4	67,1	65,8
Moyenne pour les thèmes 61 à 90	10,0	55,2	56,0	45,6

Figure 7.30 : *Ensemble des résultats sur le corpus Reuters avec les lemmes.*

Comme précédemment, l'utilisation de lemmes détériore légèrement les résultats. La Figure 7.31 qui permet de visualiser les différences pour chaque catégorie (pour les soixante premières) montre peu de différences entre les méthodes.

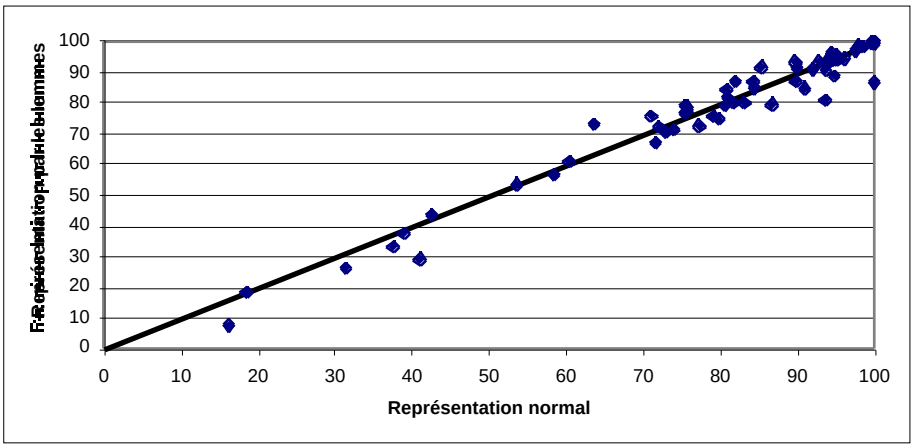

Figure 7.31 : *Comparaison des précisions moyennes non interpolées pour les soixante premiers thèmes : représentation normal ou représentation avec des lemmes.*

L'effet de la lemmatisation sur les problèmes de catégorisation de textes a été moins étudié que l'effet de substitution des mots par leur racine. [Kindermann et Leopold, 2000] ont utilisé des machines à vecteurs supports pour de la catégorisation de textes sur un corpus en langue allemande et ont comparé les performances obtenues lorsque les mots sont conservés tel quel ou lorsqu'ils sont remplacés par leur lemme. Ils observent une diminution des performances avec l'utilisation des lemmes, bien que la langue allemande soit une langue qui présente beaucoup de flexions.

Finalement l'utilisation de racines ou de lemmes n'est pas souhaitable pour nos modèles, car ces approches rendent les mots plus ambigus et finalement beaucoup d'auteurs

7.6 Mise en œuvre sur le corpus TREC-8

Ce paragraphe présente les expériences qui ont été faites pour fournir les résultats de la compétition TREC-8 [Stricker *et al.*, 2000b] au cours de l'année 1999 ; toutes les remarques faites ci-dessus ne sont pas nécessairement prises en considération, car ces résultats ont été obtenus avant la date limite du 31 août 1999. Ce paragraphe détaille le choix des paramètres qui avaient été faits pour obtenir les résultats, afin d'obtenir une référence claire pour les comparaisons. Comme on l'a précisé au chapitre 3, tous les apprentissages sont effectués avec le fichier des pertinences disponible avant la compétition et les performances sont évaluées sur la base de test en fonction du fichier des pertinences fourni après la compétition.
La synthèse et les comparaisons des résultats de la tâche de filtrage de TREC-8 sont effectuées dans [Hull et Robertson, 2000] ; une description succincte des différentes approches y est également présentée.

Nous ne reprenons pas ici la description du corpus et des données, qui a été faite au chapitre 3.

7.6.1 Les paramètres de la sélection des descripteurs

Pour chaque thème, on détermine la liste du vocabulaire spécifique grâce à la méthode exposée au chapitre 5.

La sélection de descripteurs est effectuée en construisant la matrice X nécessaire à l'orthogonalisation de Gram-Schmidt. On considère systématiquement, pour chaque thème, les 50 premiers mots de la liste du vocabulaire spécifique (cinquante descripteurs initiaux avec la terminologie utilisée jusqu'ici) et 1300 dépêches non pertinentes sélectionnées aléatoirement.
En utilisant l'algorithme d'orthogonalisation de Gram-Schmidt couplé avec le critère d'arrêt, le nombre moyen de descripteurs sélectionnés sur l'ensemble des cinquante thèmes est de vingt-cinq.

7.6.2 Apprentissage du réseau de neurone

Pour constituer la base d'apprentissage, l'ensemble des documents pertinents disponibles est pris en considération, et le nombre de documents non pertinents utilisés est choisi en fonction du nombre de documents pertinents : 2000 s'il y a plus de 30 documents pertinents et 1300 sinon.

Pour coder les composantes x_i des vecteurs d'entrées, on utilise le codage suivant :

$$x_i = \begin{cases} -1 & si\ TF_j(i) = 0 \\ \dfrac{TF_j(i)}{Log(L_j)} & si\ TF_j(i) > 0 \end{cases}$$

$TF_j(i)$ est la fréquence d'un terme i dans un texte j et L_j est la longueur de ce texte mesurée par le nombre de mots.

Le réseau de neurones est une régression logistique comme à la Figure 7.2, l'apprentissage s'effectuant en minimisant l'entropie croisée. La régularisation est faite grâce à la méthode de l'arrêt prématuré : la minimisation est faite avec une simple descente de gradient pendant quelques centaines d'itérations. Comme cet algorithme ne converge pas vers le minimum de la fonction de coût, les poids ne prennent pas de grandes valeurs et le surajsutement est évité.

7.6.3 Résultats de la compétition TREC-8

Pour la tache de routing à laquelle nous avons participé, les performances sont mesurées par la précision moyenne non interpolée : on calcule cette valeur pour chaque thème, puis la performance globale est simplement la moyenne de l'ensemble. Les thèmes ne comportant aucun document pertinent sur la base de test sont comptés avec une performance de zéro.

Sur l'ensemble des cinquante thèmes, la performance globale est de **30,7** (la performance maximale possible étant de 96,0 car deux thèmes n'ont aucun document pertinent sur le test).

Les résultats de l'ensemble des participants sont présentés à la Figure 7.32, chaque histogramme représentant un résultat. Les différents systèmes sont tous repérés par un nom qui est le nom officiel donné lors de la compétition ; notre système porte le nom **S2N2**.
Excepté pour notre système, chaque participant a présenté deux ensembles de résultats correspondant, en général, à des jeux de paramètres différents.

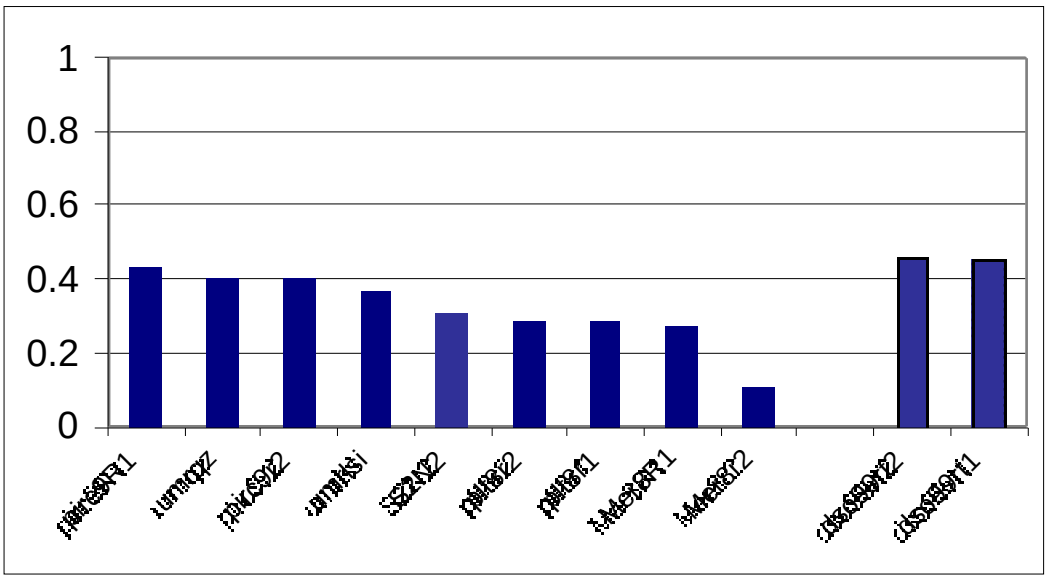

Figure 7.32 : *Moyenne des résultats obtenus sur l'ensemble des cinquante thèmes pour les participants de la tâche du routing.*

7.6.4 Revue des différents systèmes de la sous tâche de routing

Dans ce paragraphe, nous détaillons rapidement les approches utilisées par chacun des candidats, avec les références bibliographiques qui s'y rapportent.

[Ng *et al.*, 2000] ont présenté les systèmes ***dso99rt2*** et ***dso99rt1***. Leurs résultats sont présentés séparément des autres candidats, car ils ont utilisé comme descripteurs potentiels des annotations manuelles qui figuraient dans les documents du *Financial Times*. Ces annotations manuelles n'ont pas été utilisées par les autres candidats et, par conséquent, il est difficile de faire des comparaisons directes puisque leur modèle a été construit avec des informations supplémentaires par rapport aux autres.

Leurs deux systèmes reposent sur l'utilisation de l'algorithme du perceptron. La sélection des descripteurs est faite grâce à une méthode issue de l'algorithme de Rocchio amélioré [Singhal, 1998] [Schapire *et al.*, 1998], qui permet d'obtenir un classement des descripteurs par ordre de pertinence décroissante. Les cent premiers descripteurs trouvés, ainsi que les vingt meilleures paires (une paire étant définie par deux mots non vides adjacents), pour chaque thème, sont utilisés en entrée.

[Kwok *et al.*, 2000] ont présenté les systèmes ***pirc9r1*** et ***pirc9r2***. Pour le système *pirc9r1*, six profils différents sont fabriqués avec des coefficients différents. Parmi ces six profils, deux utilisent uniquement la requête, et les quatre restants utilisent la base des documents pertinents. Ces différents profils sont combinés linéairement afin d'obtenir un score pour chaque document, les coefficients affectés à chaque profil sont trouvés par un algorithme génétique pour d'optimiser la précision moyenne non interpolée.
Pour le deuxième système *pirc9r2* deux nouveaux profils sont ajoutés à la combinaison.

[Oard et Wang, 2000] ont présenté les systèmes ***umrlqz*** et ***umrlsi***. Pour leurs expériences, ils utilisent des méthodes issues de la recherche d'informations et plus particulièrement l'algorithme de Rocchio pour le système *umrlqz*. Pour le système *umrlsi*, ils essayent de faire du filtrage collaboratif grâce à la méthode *latent semantic indexing* pour trouver des structures communes à des sous-ensembles de thèmes. L'utilisation de cette deuxième méthode dégrade les résultats, peut être parce qu'il n'existe pas particulièrement de points communs entre les différents thèmes.

[MacFarlane et Robertson, 2000] ont présenté les systèmes ***plt8r1*** et ***plt8r2***. Ils ont cherché à mesurer les performances du système PLIERS qui s'appuie sur la stratégie du système Okapi de TREC 5 [Beaulieu *et al.*, 1997] : une partie de la base d'apprentissage est utilisée pour l'extraction de descripteurs et une autre partie pour la sélection. Leurs expériences ont été faites en utilisant seize ordinateurs dotés de Pentium II en parallèle qui indexent la base d'apprentissage (64.000 documents) en 5 minutes, et la base de test (140.000 documents) en 11 minutes.

[Boughanem *et al*., 2000] ont présenté les systèmes ***Mer8r1*** et ***Mer8r2*** qui utilisent un réseau de neurones pour implémenter le modèle probabiliste, couplé à des algorithmes génétiques pour trouver les paramètres d'apprentissage. Ce modèle est donc directement issu des méthodes de recherche d'informations et non pas des méthodes de catégorisation de textes.

7.6.5 Résultats obtenus après la compétition

Nous présentons dans ce paragraphe les résultats que nous avons obtenus après la compétition ; ils prennent en considération les observations faites précédemment et mettent en avant leur influence bénéfique.

Une première expérience est réalisée, pour laquelle la procédure de sélection des descripteurs n'est pas modifiée, mais les changements par rapport au système S2N2 sont les suivants :

1. Utilisation d'un terme de *weight decay* plutôt que la méthode de l'arrêt prématuré.
2. Fabrication des bases d'apprentissage en supprimant les vecteurs nuls.
3. Codage Lnu des vecteurs d'entrée (le codage Lnu a été présenté au chapitre 5).

Le modèle est toujours celui de la Figure 7.2 avec un hyperparamètre fixé à 1.

Avec ces nouveaux paramètres, la moyenne des précisions moyennes non interpolées sur les cinquante thèmes (comme précédemment, les thèmes sans document pertinent sont pris en considération dans la moyenne avec un score nul) devient **34,8** contre 30,7 précédemment.

La Figure 7.33 présente les courbes rappel-précision pour l'ensemble des thèmes, la courbe en pointillé représente le système S2N2 tandis que la courbe en trait plein est obtenue avec les nouveaux paramètres.

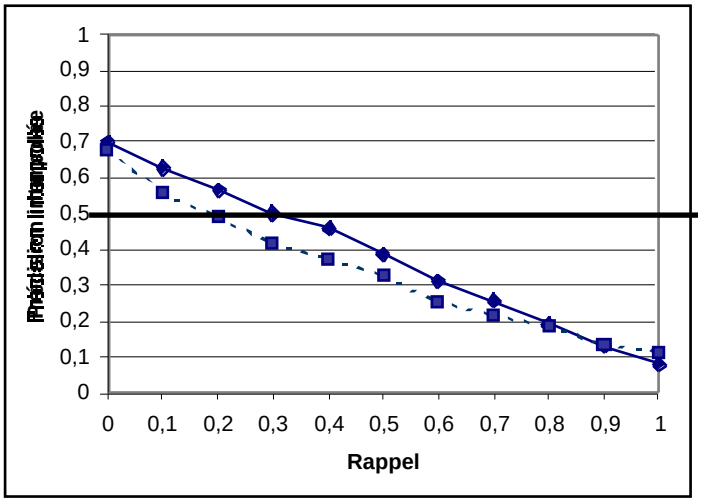

Figure 7.33 : *Courbes rappel-précision pour l'ensemble des 48 thèmes du corpus ayant des documents pertinents sur la base de test. La courbe en pointillée représente le système S2N2 présenté a TREC-8, la courbe en trait plein le nouveau système.*

L'ensemble des modifications a permis d'améliorer significativement les résultats ; cette amélioration se traduit à la fois par une amélioration de la moyenne des précisions moyennes non interpolées, et par le fait que la courbe rappel-précision de l'ensemble des thèmes est systématiquement au-dessus (sauf pour le rappel de 1).

Ces résultats confirment ce qui a été vu au paragraphe 7.3 : la suppression des vecteurs nuls améliore les performances sur la base de test. D'autre part, le codage Lnu proposé dans [Singhal, 1996] est un codage plus performant que notre codage initial. Et de même, la méthode du *weight decay* s'est avérée supérieure à la méthode de l'arrêt prématuré.

Influence de la valeur de l'hyperparamètre

Pour l'expérience précédente, l'hyperparamètre a été fixé à 1, pour l'ensemble des thèmes. En continuant d'utiliser une valeur identique pour l'ensemble des thèmes, il peut être intéressant de voir l'influence globale que peut avoir cette valeur ; plusieurs valeurs de l'hyperparamètre α ont été testées : 0, 0,5, 1, 5, 10. La valeur nulle de l'hyperparamètre correspond à une fonction de coût sans terme de régularisation.

La Figure 7.34 donne les performances obtenues sur l'ensemble des thèmes en fonction de la valeur de l'hyperparamètre et la Figure 7.35 montre l'évolution des courbes rappel-précision interpolée

	α = 0	α = 0,5	α = 1,0	α = 5,0	α = 10,0
Performances (UAP)	24,0	35,6	34,8	32,6	31,7

Figure 7.34 : *Comparaisons des performances en fonction de la valeur de l'hyperparamètre. La performance est la moyenne des précisions moyennes non interpolées sur les cinquante thèmes.*

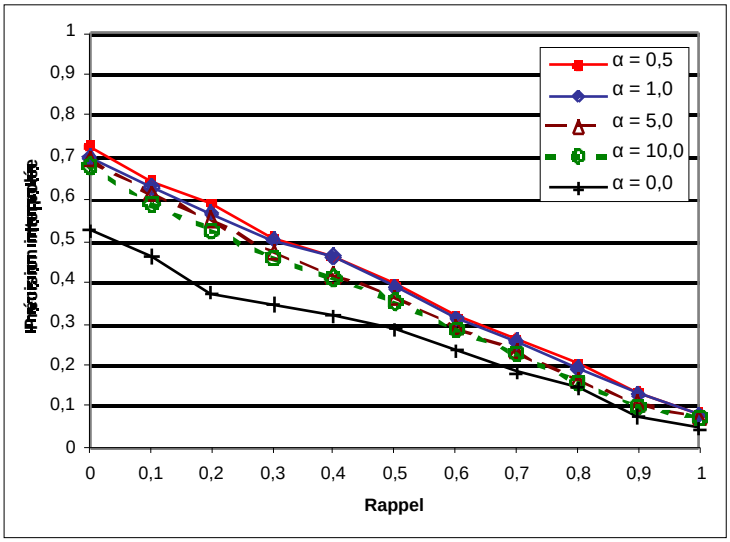

Figure 7.35 : *Courbes rappel-précision pour l'ensemble des thèmes en fonction de la valeur de l'hyperparamètre.*

À partir de l'ensemble des résultats obtenus sur TREC-8, il est possible de tirer plusieurs conclusions :

- Les résultats obtenus sans aucune régularisation sont nettement inférieurs à tous les autres, quelle que soit la valeur de l'hyperparamètre. Par conséquent, il est indispensable d'utiliser une méthode de régularisation lors de l'apprentissage.
- Les performances moyennes ne semblent pas très sensibles à la valeur de l'hyperparamètre ; les performances obtenues avec les valeurs 0,5 ou 1 sont très proches et les performances obtenues avec les valeurs 5 ou 10 sont légèrement inférieures. Il vaut donc mieux choisir des valeurs pas trop élevées pour les hyperparamètres, car les performances moyennes se dégradent légèrement lorsque cette valeur augmente.

7.7 Conclusion

Ce chapitre a permis d'introduire un modèle de filtrage, compétitif par rapport à d'autres méthodes ; dont l'efficacité a été montrée sur deux corpus différents : le corpus Reuters et le corpus de la tache de routing de TREC-8.

Les bonnes performances reposent principalement sur deux points : d'une part une sélection de descripteurs efficace et, d'autre part, l'utilisation d'un terme de *weight decay* lors de l'apprentissage. Nos modèles ont en plus la particularité d'utiliser très peu de descripteurs pour la représentation des textes : 17 en moyenne sur le corpus Reuters et 25 sur le corpus TREC-8.
Il faut noter que si l'ajout d'un terme de *weight decay* s'est révélé essentiel, les performances ne sont pas extrêmement sensibles à la valeur de l'hyperparamètre tant que celle-ci n'est pas nulle. Cependant, la méthode de détermination de l'hyperparamètre par l'approche bayesienne ne s'est pas révélée très efficace. [Wiener, 1993] qui utilise également un terme de régularisation lors des apprentissages des réseaux de neurones a choisi, lui aussi, une valeur constante plutôt que de faire varier l'hyperparamètre pendant l'apprentissage.

Finalement nos résultats vont dans le sens des résultats cités au chapitre 2 sur les approches neuronales [Wiener *et al.*, 1995] [Schütze *et al.*, 1995] : l'ajout de neurones cachés n'améliore pas les résultats et l'utilisation d'une méthode de régularisation est indispensable même avec l'architecture la plus simple.

Si les modèles obtenus sont performants, ils sont loin d'être parfaits, et il semble possible de les améliorer. Cependant ni le nombre de neurones cachés, ni les paramètres d'apprentissage (notamment la valeur de l'hyperparamètre) ne semblent être en mesure d'apporter des améliorations significatives ; de plus, l'étude du chapitre 5 sur les différentes sélections de descripteurs a montré que ces méthodes étaient équivalentes.
Pour progresser, il est nécessaire d'améliorer la qualité de représentation des textes et notamment d'inclure plus d'informations dans cette représentation, afin de définir ensuite une

architecture neuronale adéquate. C'est l'objectif de l'approche originale qui est présentée dans le chapitre suivant.