Chapitre 8 Utilisation du contexte local des mots

Les expériences du chapitre 7 ont montré que les meilleures performances étaient obtenues avec un séparateur linéaire, c'est-à-dire sans neurone caché. Les résultats d'autres auteurs confirment que ce résultat n'est pas dû spécifiquement à la méthode de sélection de descripteurs employée, mais semble être intrinsèque à la représentation des textes en sac de mots. Améliorer la représentation des textes, en enrichissant l'information qu'elle contient, peut conduire à l'utilisation de classifieurs plus complexes. La représentation en sac de mots ne tient compte ni de l'ordre, ni de la distance entre les mots. Dans ce chapitre, nous proposons une représentation des textes qui tient compte du contexte local des mots pour les désambiguïser ; l'architecture neuronale est modifiée pour tenir compte de cette nouvelle représentation.

Ce modèle a été appliqué sur les corpus Reuters et TREC-8 et a également servi pour notre participation à la sous-tâche de routing de TREC-9.

8.1 L'Ambiguïté dans la recherche de textes

8.1.1 Exemples d'ambiguïté sémantiques

La liste du vocabulaire spécifique du thème d'échanges de participations comprend le mot *participation* qui, dans un contexte économique, a un sens précis comme l'illustre la phrase suivante :

- Dexia a pris une *participation* de 35,3 % dans le capital de la banque italienne.

Or, ce mot peut être employé dans un contexte différent comme dans l'expression suivante :

- la *participation* des communistes au gouvernement.

Dans ces deux exemples, les termes proches permettent de déterminer exactement le sens du mot *participation*, car sa présence dans une phrase ne suffit pas à caractériser le concept d'échange de participations entre deux entreprises. Dans le premier exemple, la présence du

mot *capital* précise ce concept, alors que dans le deuxième exemple, la présence du mot *gouvernement* indique que le concept des entreprises est absent.

Cet exemple simple montre que la présence de mots dans le voisinage immédiat de certains mots-clefs peut renforcer, ou annihiler, le concept que le mot est susceptible de représenter.

De même, certains mots associés sous forme de paires peuvent exprimer des concepts très précis, alors que le sens de chacun des mots peut être beaucoup plus vague. Considérons, par exemple, les trois concepts suivants, construits à partir du mot *droits* :

- droits de vote
- droits de douane
- droits de l'homme

Ils définissent des concepts importants à reconnaître. Ici la proximité des mots est primordiale : ce n'est pas parce que *droits* et *vote* figurent dans le même texte que le concept de *droit de vote* est présent. La représentation en sac de mots ne reconnaît pas, de manière systématique, la présence de ces associations.

Par exemple, dans les deux extraits de phrases suivantes, provenant du corpus Reuters, deux sens différents du mot *interest* sont présentés :

- (…) to the decline in the value of the U.S dollar by raising *interest* rates (…).
- Soviet officials said foreign businessmen are expressing strong *interest* in establishing joint enterprises in the Soviet Union (…).

Dans la première phrase, le sens du mot *interest* est clairement précisé par la présence de *rates* juste à côté pour former l'expression *interest rates* (taux d'intérêt) et dans une moindre mesure par la présence du verbe *raising*. Dans la deuxième phrase, l'absence de ces termes indique que le sens n'est pas celui des taux d'intérêt.

8.1.2 Autres travaux sur la désambiguïsation et la recherche d'informations

Plusieurs auteurs ont cherché à savoir si les méthodes de désambiguïsation pouvaient avoir un impact positif sur les systèmes de recherche d'informations et, plus précisément, quelles méthodes de désambiguïsation pouvaient y parvenir. On peut distinguer deux approches générales pour effectuer cette désambiguïsation dans le cadre de la recherche d'informations.

La première approche suppose qu'il existe un nombre fini de sens pour un mot, et repose sur l'utilisation de dictionnaires. Par exemple, [Voorhees, 1993] utilise WordNet[1] [Miller *et al.*, 1990] pour la désambiguïsation de mots en langue anglaise, mais observe une décroissance des performances. Le problème de l'utilisation d'un tel dictionnaire est qu'il suppose qu'il existe un nombre fini de sens pour un mot donné, et d'autre part, il ne couvre pas nécessairement toute l'étendue de vocabulaire. De plus, certaines nuances proposées peuvent posséder un sens linguistique, mais n'être pas nécessairement utiles pour la recherche d'informations. Enfin, certaines nuances peuvent dépendre du domaine que l'on cherche à filtrer : par exemple, dans les deux phrases suivantes qui utilisent le mot *capital* :

- Il détient le *capital* de la société X.
- La société X propose du *capital*-risque.

Le mot *capital* a des sens très proches, mais pour certains filtres, il peut être utile de les différencier.

La deuxième approche repose sur l'utilisation du corpus, et plus précisément, sur une étude du contexte des mots, pour effectuer une désambiguïsation utilisable par les systèmes de recherche d'informations.

L'utilisation la plus simple du contexte consiste à utiliser des paires de mots définies comme étant deux mots non vides adjacents. [Singhal, 1998] a obtenu, par cette méthode, des améliorations de performances sur le corpus de TREC-6 et [Ng *et al.*, 2000] ont également utilisé cette définition des paires pour TREC-8. En revanche [Dumais *et al.*, 1998] observe une diminution des performances de leur système lorsqu'ils considèrent l'utilisation de paires en entrée de leurs machines à vecteurs supports. Le problème qui se pose avec les paires est évidemment celui de l'ordre des mots : par exemple si l'on cherche les documents à propos de "*car insurance rates*", et que la paire *insurance_rates* est utilisée, les documents qui parlent de "*rates for insurance car*" ne vont pas être sélectionnés.

[1] http://www.cogsci.princeton.edu/~wn

D'autres auteurs utilisent une notion de contexte plus large pour effectuer la désambiguïsation. Par exemple, [Cohen et Singer, 1996] ont testé deux notions différentes de contextes. Pour leur système RIPPER, construit à base de règles, le contexte d'un mot est défini comme une liste de mots (généralement de faible taille) qui doivent apparaître en même temps dans n'importe quel ordre et n'importe où dans le texte. Pour leur deuxième système, appelé *sleeping experts*, le contexte est constitué de plusieurs mots ordonnés et proches les uns des autres. Les deux approches, bien que différentes, donnent de bons résultats et montrent que l'utilisation du contexte peut améliorer la qualité des filtres.
Pour [Yarowski, 1995] un mot a un sens principal très corrélé aux mots figurant juste à côté. Pour mettre en œuvre cette approche, [Jing et Tzoukermann, 1999] [Schütze et Pedersen, 1995] définissent des vecteurs de contextes pour désambiguïser les mots et obtiennent une amélioration des performances.

Notre approche est proche de celle développée par [Jing et Tzoukermann, 1999], et repose sur l'utilisation du corpus pour définir le contexte usuel d'un mot. Pour chaque mot, le vecteur de contexte de chaque mot est constitué des cinq mots qui le précèdent et des cinq mots qui le suivent.

8.2 Détermination automatique de vecteurs de contexte

8.2.1 Définition du contexte local d'un mot

Dans la représentation en sac de mots, les descripteurs utilisés pour discriminer les textes pertinents des textes non pertinents sont des mots simples pris séparément les uns des autres. Or, le contexte d'un mot dans le sous-ensemble des textes pertinents est différent du contexte de ce mot dans le sous-ensemble des textes non pertinents. Cette différence peut donc être utilisée pour la discrimination.

Dans toute la suite, le contexte d'un mot est défini par une fenêtre de dix mots : les cinq mots qui le précédent et les cinq mots qui le suivent.
Par exemple, dans la phrase suivante :

La société de parfumerie Marionnaud **détient** désormais 292.157 actions, soit 8,09 % du capital.

Le contexte du mot **détient** est défini par un vecteur dont les composantes sont les occurrences des mots (*actions, capital, de, désormais, du, la, marionnaud, parfumerie, société, soit*). Les chiffres ne sont pas pris en considération, et l'ordre des mots à l'intérieur de la fenêtre n'a pas d'importance.

Si l'on considère un ensemble de textes, il est possible d'additionner tous les vecteurs de contextes trouvés pour un mot donné, et de classer ensuite tous ces contextes par ordre de d'occurrence décroissante. Mais, dans ce cas, les mots qui apparaissent avec la plus grande occurrence sont les mots les plus fréquents du corpus qui n'apportent pas d'information.

Pour définir un contexte utile, on s'appuie sur la méthode de détermination du vocabulaire spécifique présentée au chapitre 5, qui permet d'éliminer automatiquement les mots fréquents et les mots rares.

Si l'on note *TF(m, t)*, la fréquence d'un mot *m* dans un texte *t*, et *CF(m)* la fréquence de ce mot sur l'ensemble du corpus, on calcule, pour chaque mot d'un texte *t*, le rapport :

$$R\,(m,t) = \frac{TF\,(m,t)}{CF\,(m)}$$

Les mots du texte sont classés par ordre décroissant de la valeur de ce rapport, et la deuxième moitié de la liste est supprimée. On obtient, pour chaque texte, une liste *L(t)* de mots parmi lesquels les mots fréquents ont été éliminés.

Pour un mot donné, on ne tient compte d'un contexte que s'il figure dans la liste *L(t)*. Grâce à cette méthode, les mots fréquents ne sont pas pris en considération. En additionnant tous les vecteurs trouvés, et en classant les mots trouvés par ordre décroissant, les mots rares se retrouvent en fin de liste comme lors de la détermination du vocabulaire spécifique.

8.2.2 Exemples de vecteurs de contexte local

Les exemples ci-dessous montrent des contextes trouvés par cette méthode, pour des thèmes issus de corpus différents. Pour chacun de ces thèmes, le contexte de certains mots issus du vocabulaire spécifique est présenté. Le contexte de ces mots est déterminé sur un ensemble de documents pertinents pour le thème étudié, et sur un sous-ensemble de documents non

pertinents, pour mettre en évidence les différences. Par analogie avec des définitions déjà utilisées, le contexte trouvé à partir de l'ensemble des documents pertinents de la base d'apprentissage est appelé **contexte positif,** celui trouvé à partir des documents non pertinents est appelé **contexte négatif**.

La Figure 8.1 montre le contexte obtenu pour certains mots du vocabulaire spécifique du thème *participation*. Les mots dont on étudie le contexte figurent en gras, et sont suivis de leur contexte (s'il existe, c'est-à-dire s'il apparaît dans deux documents différents au moins). La colonne de gauche est le contexte positif, celle de droite présente le contexte négatif.

En comparant chacune des colonnes, on s'aperçoit aisément que le contexte dans lequel apparaissent les mots est souvent très différent selon le sous-ensemble de textes considérés (pertinents ou non pertinents). Dans le premier exemple, la présence du mot *capital* dans l'environnement immédiat du verbe *détient* précise le sens de ce verbe et d'une certaine manière le désambiguïse. On peut noter également la présence d'adverbes comme *désormais* ou *actuellement*, car, pour rendre compte des échanges de participations, ils apparaissent très fréquemment dans des tournures de phrases telles que :

- Après cette opération, la société X *détient désormais* 5 % du capital.
- La société X *détient actuellement* 5% du capital.

Par conséquent, il semble que, sur le corpus de l'AFP que nous utilisons, les adverbes *désormais* et *actuellement* précise le contexte dans lequel est utilisé le verbe détenir.

Contexte positif	**Contexte négatif**
détient	**détient**
capital	titres
actions	portefeuille
désormais	
droits	
participation	
société	
holding	
actuellement	
vote	
parts	
participation	**participation**
prise	sommet
capital	doutes
majoritaire	élève
prendre	éventuelle
minoritaire	président
détient	sérieux
céder	importante
pris	résultat

	sa		faveur
	porter		étrangère
capital		**capital**	
	détient		augmentation
	droits		risque
	augmentation		fonds
	participation		étranger
	actions		propres
	vote		investi
	acquérir		banques
	détenu		action
	représentant		ordinaire
	société		compte

Figure 8.1 : *Contexte spécifique des mots pour le thème participation.*

La Figure 8.2 est obtenue de la même manière pour le thème *partenariat* qui traite des accords de coopération et des partenariats entre entreprise.

Dans ce cas également, les mots issus du vocabulaire spécifique apparaissent dans des contextes différents, et le contexte permet de distinguer un contexte économique d'un contexte de politique de coopération internationale.

Contexte positif		**Contexte négatif**	
partenariat		**partenariat**	
	stratégique		relations
	commercial		signé
	conclu		contrat
	signature		paix
	signé		transatlantique
	accords		
	industriel		
	accord		
	global		
	renforcer		
coopération		**coopération**	
	accord		internationale
	accords		accords
	signé		développement
	domaine		franco
	conclu		étroite
	industrielle		économiques
	commerciale		renforcer
	renforcer		bilatérale
	domaines		matière
	franco		domaines
alliance		**alliance**	
	stratégique		agricole
	commerciale		atlantique
	annoncée		propos
	groupes		and
	compagnies		militaire
	possibilité		éditrice
	capitalistique		fusion
	accord		opposition
	vue		tourisme
	conclure		recours

Figure 8.2 : *Contextes spécifiques des mots pour le thème partenariat.*

La Figure 8.3 présente les résultats obtenus avec le thème *interest* du corpus Reuters. Le contexte du mot *rate* permet, par exemple, de faire la différence lorsque ce mot apparaît pour parler des taux de changes avec les expressions *floating rate* ou *exchange rate,* contrairement au cas où il apparaît dans un contexte de taux d'intérêt avec l'expression *prime rate.*

Enfin, la Figure 8.4 présente le contexte spécifique obtenu pour le thème 351 de TREC-8 (*Falkland petroleum exploration*) pour *islands* et *oil.* L'étude de ce contexte montre bien que *oil* ne doit être un indicateur que s'il est suivi de notions indiquant que l'on parle des îles malouines (*islands, malvinas*, ou *argentina*).

	Contexte positif		Contexte négatif
money		**money**	
	stg		supply
	rates		m
	k		growth
	market		stg
	call		k
	given		england
	further		liquidity
	supply		broad
	assistance		assistance
	rate		given
rates		**rates**	
	base		interest
	interest		levels
	lending		freight
	cut		current
	money		interbank
	point		stability
	k		exchange
	rates		points
	short		lower
	term		short
rate		**rate**	
	prime		floating
	cut		fixed
	base		dollar
	effective		rate
	lending		exchange
	discount		yen
	maturity		growth
	funds		discount
	point		variable
	interest		inflation

Figure 8.3 : *Contextes spécifiques des mots pour le thème interest du corpus Reuters.*

	Contexte positif		Contexte négatif
islands		**islands**	
	malvinas		highlands
	falkland		enterprise
	oil		channel
	sovereignty		china
	argentina		spratly
	argentine		cayman4
	aires		madeira
	buenos		shetland
	islands		claim
	tella		development
oil		**oil**	
	islands		gas
	malvinas		crude
	argentina		production
	oil		barrels
	exploration		exploration
	exploitation		bn
	existence		fields
	tella		africa
	ypf		saudi
	argentine		m

Figure 8.4 : *Contextes spécifiques des mots pour le thème 351 du corpus TREC-8 (Falkland petroleum exploration).*

8.3 Modèle neuronal avec contexte

8.3.1 Architecture prenant en considération le contexte

Nous venons de voir que le contexte devait, en fait, servir à renforcer ou à diminuer l'influence d'un mot. Les entrées de la régression logistique ne sont plus de simples descripteurs comme dans le cas de la représentation en sacs de mots, mais les sorties de neurones dont les entrées sont représentées sur la Figure 8.5. Le *mot principal* est issu de la méthode de sélection de descripteurs exposée au chapitre 7, et le contexte est déterminé par la méthode exposée dans le paragraphe précédent.

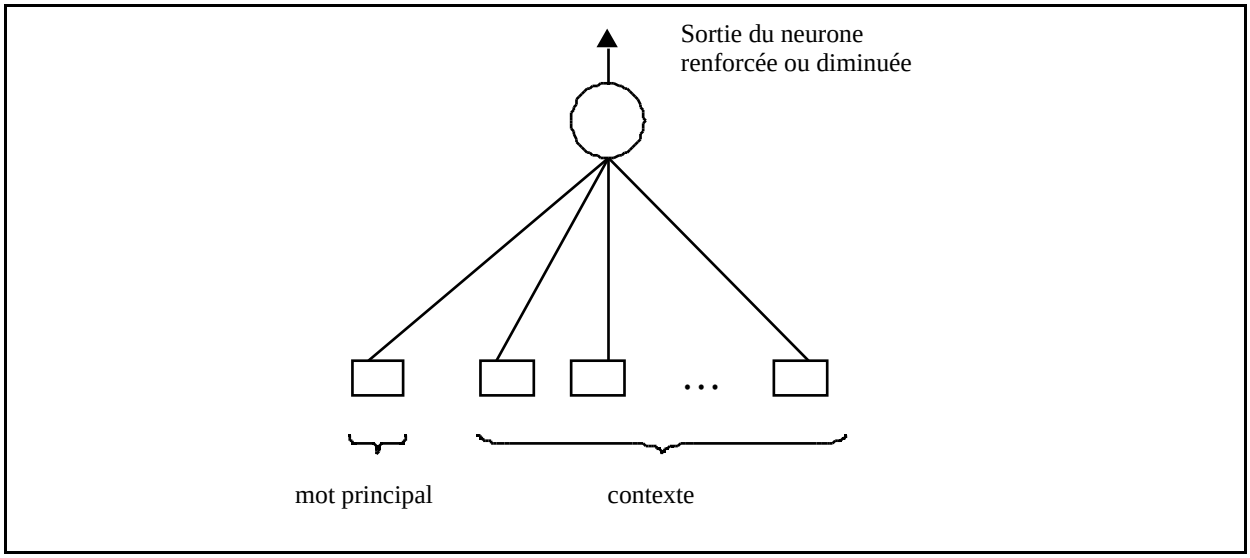

Figure 8.5 : *Unité de base de la nouvelle architecture neuronale (le biais n'est pas dessiné).*

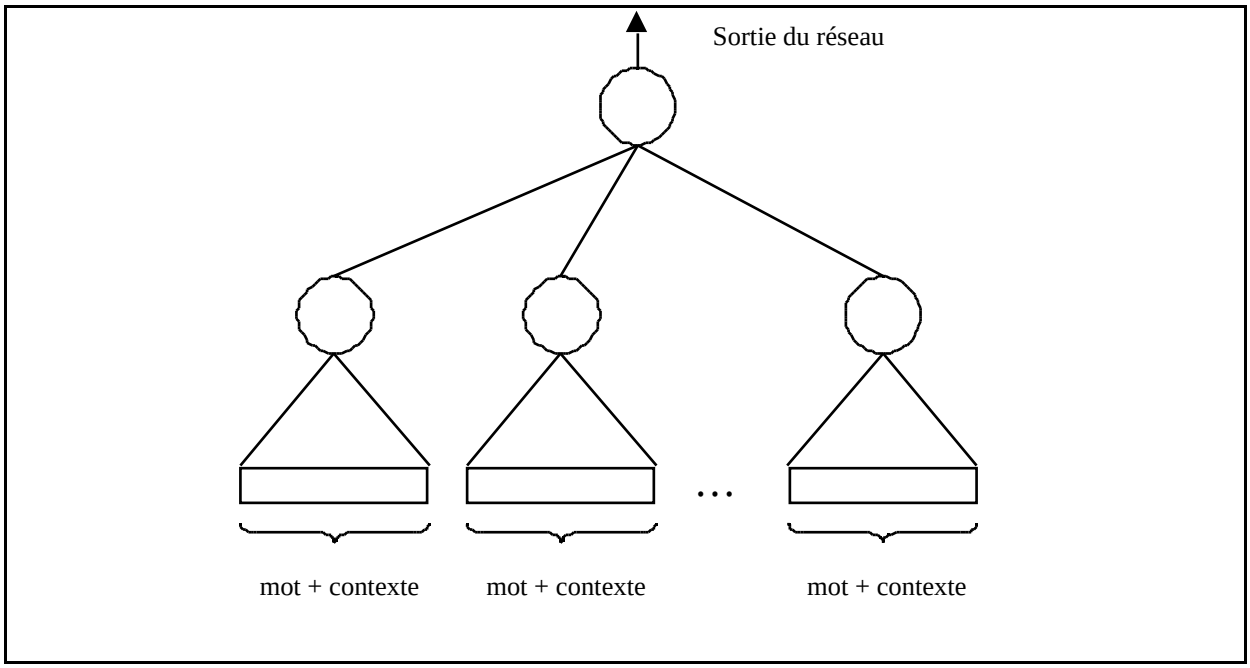

Figure 8.6 : *Architecture neuronale pour tenir compte du contexte.*

Le réseau de neurones complet est décrit par la Figure 8.6 : il prend en considération le contexte de chaque mot, et ce qui était considéré précédemment comme un descripteur d'entrée est un neurone caché dans cette nouvelle architecture.

Cette architecture est proche de l'architecture utilisée précédemment, la première couche de neurones permettant de préciser l'influence d'un descripteur particulier en fonction de son contexte.

8.3.2 Combien de poids dans la nouvelle architecture ?

Cette architecture fait intervenir plus de paramètres ajustables que la régression logistique du chapitre 7. Si, par exemple, le nombre de descripteurs sélectionnés était de 30 alors le modèle comprenait 31 paramètres (avec le biais du neurone de sortie). Si, chaque descripteur précédent est précisé par cinq contextes possibles, alors le réseau de neurones décrit à la Figure 8.5 contient 30 neurones cachés, chacun étant lié à sept entrées (le mot principal, les cinq contextes et le biais) ; le nombre de poids à déterminer lors de l'apprentissage est :

$$7*30+30+1 = 241$$

Le nombre de poids augmente considérablement par rapport à l'architecture précédente, mais la taille des bases d'apprentissage n'a évidemment pas augmenté : il faut vérifier qu'il est effectivement possible de mener l'apprentissage correctement.

8.3.3 La régularisation mise en œuvre par la méthode du weight decay

Les expériences des chapitres précédents ont montré que l'utilisation d'un terme de *weight decay* était indispensable pour obtenir de bonnes performances. Cette remarque est encore plus vraie avec cette nouvelle architecture, en raison de l'augmentation du nombre des poids.

L'architecture de la Figure 8.6 est une architecture de perceptron multi-couche partiellement connectée : il est donc nécessaire d'utiliser trois hyperparamètres α_1, α_2 et α_3 pour le terme de *weight decay* et de minimiser la fonction de coût suivante :

$$EC(w) + \frac{\alpha_0}{2} \sum w_i^2 + \frac{\alpha_1}{2} \sum w_i^2 + \frac{\alpha_2}{2} \sum w_i^2 + \frac{\alpha_3}{2} \sum w_i^2$$

où $EC(w)$ est l'entropie croisée, W_0 représente l'ensemble des poids des connexions reliant les biais aux neurones cachés, W_1 représente l'ensemble des poids des connexions reliant les entrées aux neurones cachés et W_3 représente l'ensemble des poids liés des connexions du neurone de sortie y compris le biais du neurone de sortie.

8.4 Détermination des valeurs des hyperparamètres

Afin de tester l'influence des valeurs des hyperparamètres sur l'apprentissage de l'architecture neuronale présentée sur la Figure 8.6, trois approches différentes sont mises en œuvre :

- la première consiste à fixer les valeurs *a priori* et à conserver ces valeurs fixes pendant l'apprentissage,
- la deuxième consiste à faire varier les hyperparamètres pendant l'apprentissage par la méthode de maximisation issue du formalisme de l'approche bayesienne décrit au chapitre 6**,**
- la dernière consiste à faire varier les hyperparamètres selon la méthode d'intégration issue du formalise de l'approche bayesienne décrit au chapitre 6.

Ces expériences sont effectuées sur les 90 catégories du corpus Reuters ; pour chaque catégorie, le mot principal de l'architecture de la Figure 8.5 correspond aux résultats de la sélection de descripteurs du chapitre 7, et pour chacun de ces mots, les cinq premiers contextes positifs sont ajoutés.

Trois hyperparamètres sont utilisés selon la répartition décrite au paragraphe 8.3.3.

8.4.1 Hyperparamètres constants

Pour cette expérience, les trois hyperparamètres sont constants durant tout l'apprentissage ; on cherche à étudier l'impact du choix des valeurs sur les résultats. Plusieurs expériences ont montré que la valeur de l'hyperparamètre α_0 concernant le biais des neurones cachés avait peu d'importance : cet hyperparamètre est fixé à 0,001 pour toutes les expériences suivantes.

Les deux hyperparamètres α_1 et α_2 varient dans l'intervalle [0 ; 6] par pas de 0,25. Pour chaque valeur du couple (α_1, α_2), l'apprentissage est effectué, et la précision moyenne non interpolée (UAP) ainsi que la valeur de *F* optimale sont calculées sur la base de test. Le couple (0, 0) correspond à un apprentissage effectué sans régularisation (excepté pour le biais des neurones cachés).

Les macro-moyennes sont calculées par sous-groupe de catégories afin de relier les variations aux nombres d'exemples pertinents présents dans la base d'apprentissage.

Les résultats sont présentés à la Figure 8.7 pour chaque sous-groupe. La colonne de gauche montre l'évolution de la précision moyenne non interpolée et la colonne de droite l'évolution de la valeur de *F*. L'axe des X est la valeur de l'hyperparamètre α_1 (connexions reliant les entrées aux neurones cachés) et l'axe des Y la valeur de l'hyperparamètre α_2 (connexions du neurone de sortie). L'échelle selon l'axe Z varie selon le groupe de catégories considéré pour mettre en évidence les variations.

Quel que soit le sous-groupe de catégories considéré, l'ensemble des résultats prouve qu'il est indispensable d'utiliser une méthode de régularisation lors de l'apprentissage, car, lorsque l'un des hyperparamètres a une valeur nulle, les performances sont nettement détériorées. Pour la précision moyenne non interpolée, les performances obtenues avec le couple (0, 0) sont systématiquement les plus mauvaises.

Pour l'ensemble des catégories, la précision moyenne non interpolée est peu affectée par la valeur des hyperparamètres à partir du moment où ils ne sont pas nuls. Pour les soixante premières catégories, elle diminue légèrement lorsque les deux hyperparamètres prennent des valeurs supérieures à deux. Pour les trente dernières catégories (catégories 61 à 90), le comportement est légèrement différent de celui observé pour les autres sous-groupes, car les performances ne décroissent pas lorsque les hyperparamètres ont des valeurs élevées.

L'évolution de la valeur de *F* avec les hyperparamètres montre que les performances ont tendance à décroître lorsque les valeurs des deux hyperparamètres augmentent ; le phénomène est d'autant plus prononcé que le nombre de documents pertinents diminue sur la base d'apprentissage.

Pour les catégories ayant peu de documents pertinents, la valeur de *F* diminue significativement lorsque les valeurs des deux hyperparamètres augmentent, alors que la précision moyenne non interpolée reste à un niveau comparable. Ce comportement différent des deux mesures a déjà été rencontré au chapitre 6 : la sortie du réseau de neurones prend des valeurs très faibles, et, si le classement est toujours possible, l'utilisation d'un seuil ne permet plus de séparer les documents.

Pour des valeurs égales des hyperparamètres, plus la taille de la base d'apprentissage est faible, moins le terme d'entropie croisée a d'importance dans la fonction de coût total, et plus les termes de *weight decay* sont importants. Par conséquent, les poids tendent rapidement vers zéro lors de l'apprentissage. Néanmoins, l'apprentissage s'effectue toujours, et les documents pertinents ont globalement une probabilité de pertinence plus élevée que les documents non pertinents, mais du fait de la faible valeur des poids, ces probabilités sont proches de zéro.

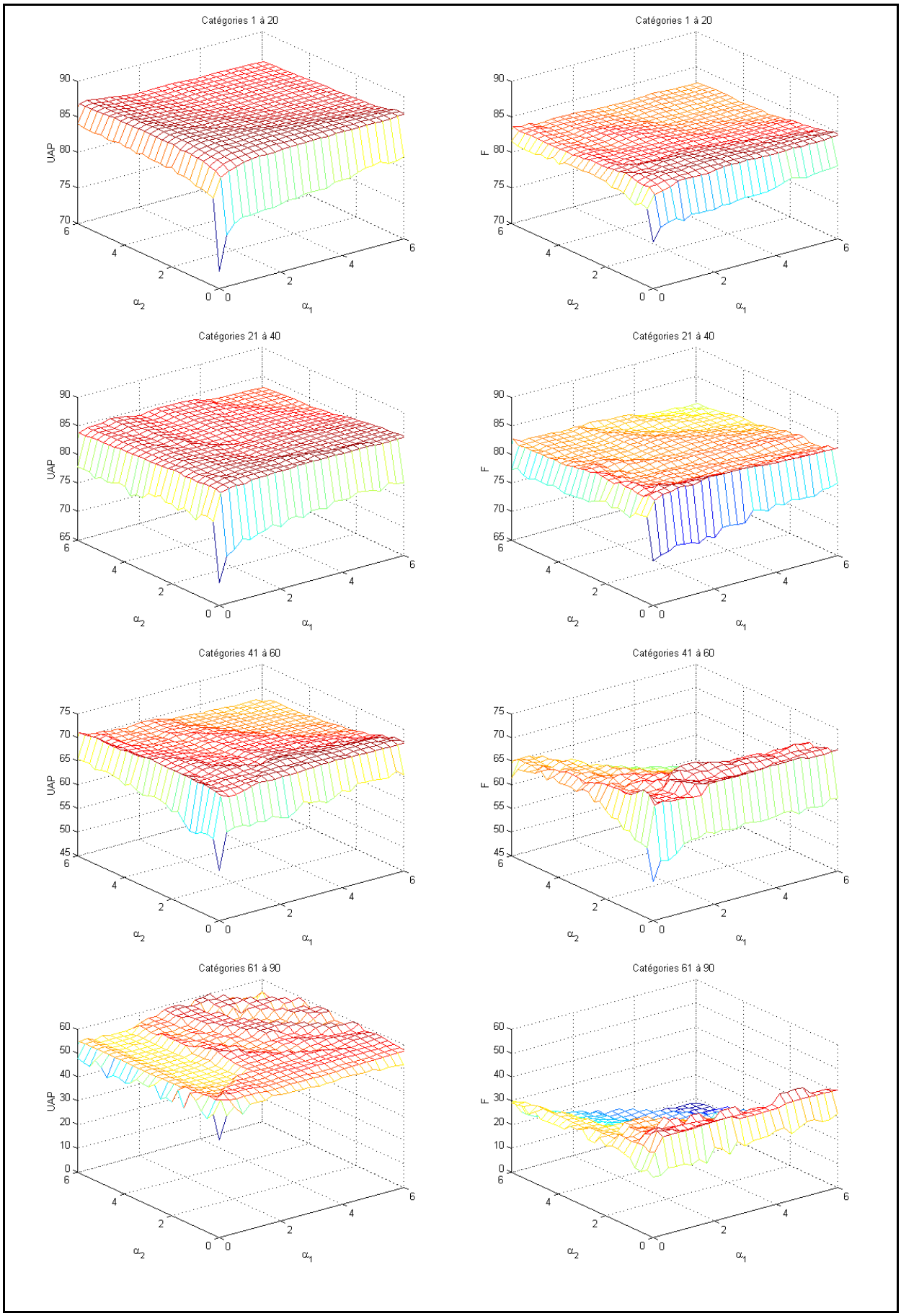
Catégories 1 à 20
UAP
Catégories 1 à 20
F
Catégories 21 à 40
UAP
Catégories 21 à 40
F
Catégories 41 à 60
UAP
Catégories 41 à 60
F
Catégories 61 à 90
UAP
Catégories 61 à 90
F
α1
α2

Figure 8.7 : *Évolution des performances (F et UAP)selon la valeur des hyperparamètres.*

En conclusion, les expériences montrent qu'il est possible de choisir des constantes pour les valeurs des hyperparamètres, car les performances ne sont pas extrêmement sensibles à ces valeurs à partir du moment où elles ne sont pas trop élevées. Pour l'hyperparamètre lié aux connexions du neurone de sortie, une valeur faible semble préférable alors que, pour l'hyperparamètre lié aux connexions entre les entrées et les neurones cachés, il est possible de choisir des valeurs légèrement supérieures.

Pour les comparaisons ultérieures, la Figure 8.8 récapitule les résultats obtenus sur l'ensemble des thèmes avec $\alpha_0 = 0{,}001$, $\alpha_1 = 1{,}0$ et $\alpha_2 = 0{,}5$.

	UAP	*F*
Moyenne sur l'ensemble des thèmes	72,28	65,6
Moyenne pour les thèmes 1 à 10	90,30	86,27
Moyenne pour les thèmes 11 à 40	84,72	83,32
Moyenne pour les thèmes 41 à 60	70,45	67,99
Moyenne pour les thèmes 61 à 90	55,52	39,76

Figure 8.8 : *Résultats obtenus avec $\alpha_0 = 0{,}001$, $\alpha_1 = 1{,}0$ et $\alpha_2 = 0{,}5$.*

8.4.2 Hyperparamètres optimisés : méthode d'intégration

Dans le paragraphe précédent, les hyperparamètres étaient fixes durant l'apprentissage. Pour l'expérience décrite ici, les hyperparamètres évoluent durant l'apprentissage, et sont déterminés par la méthode d'intégration issue de l'approche bayesienne expliquée au chapitre 6.

Dans cette approche, chaque hyperparamètre est estimé régulièrement pendant l'apprentissage selon la formule :

$$\alpha_k = \frac{\gamma}{\cdot}$$

p_k est le nombre de poids concernés par l'hyperparamètre α_k.

Pour cette expérience α_0 est fixé à 0,001 comme précédemment et n'évolue pas pendant l'apprentissage. La valeur initiale de α_1 est 0,5 et la valeur initiale de α_2 est 1,0 ; ces deux hyperparamètres sont calculés régulièrement pendant l'apprentissage.

Les résultats obtenus sont présentés à la Figure 8.9 et doivent être comparés à ceux obtenus à la Figure 8.8.

	UAP	*F*
Moyenne sur l'ensemble des thèmes	42,85	31,94
Moyenne pour les thèmes 1 à 10	89,15	84,13
Moyenne pour les thèmes 11 à 40	75,08	66,66
Moyenne pour les thèmes 41 à 60	20,35	2,78
Moyenne pour les thèmes 61 à 90	10,97	0,14

Figure 8.9 : *Résultats obtenus avec les hyperparamètres déterminés par la méthode d'intégration.*

Les résultats sont nettement inférieurs à ceux obtenus précédemment, et l'écart est d'autant plus grand que le nombre de documents pertinents sur la base d'apprentissage est faible.

En fait, avec cette approche, l'hyperparamètre α_1 tend systématiquement vers des valeurs très élevées (de l'ordre de la centaine), et par conséquent les poids liés aux entrées tendent très rapidement vers des valeurs faibles.

Les résultats dépendent de l'initialisation des hyperparamètres, mais, quelles que soient les valeurs essayées, le comportement reste identique et les performances décevantes.

8.4.3 Hyperparamètres optimisés : méthode de maximisation

La méthode de maximisation des hyperparamètres issue de l'approche bayesienne a également été testée sur ce problème. La théorie de cette approche a été détaillée au chapitre 6.

Rappelons les résultats :

$$\alpha_k = \frac{\quad}{\quad}$$

avec

$$\left(\eta_j - \left(\quad \right) \right)$$

où $\{\eta_j\}$ représente l'ensemble des valeurs propres du hessien A de la fonction de coût régularisée, V est la matrice des vecteurs propres, et I_k est la matrice ne contenant que des valeurs nulles sauf sur les éléments de la diagonale liés au groupe de poids gouvernés par l'hyperparamètre α_k où la valeur est 1.

Le paramètre γ_k peut également être calculé par la formule suivante, équivalente à la précédente :

$$\gamma_k = k - \alpha_k \mathrm{Tr}_k\left[A^{-1}\right]$$

où $\mathrm{Tr}_k\left[A^{-1}\right]$ est la trace ne portant que sur les éléments gouvernés par l'hyperparamètre α_k.

Par rapport à la méthode de maximisation, le paramètre p_k est remplacé par le paramètre γ_k dont la valeur est inférieure ou égale à p_k : on peut donc espérer corriger le défaut de la méthode précédente, qui conduisait à une valeur de l'hyperparamètre trop élevée et faisait tendre les poids vers zéro.

En pratique, comme la fonction de coût régularisée s'écrit :

$$I(w) = E(w) + \alpha \cdot E^0(w) \qquad (\quad) \qquad (\quad)$$

$E_c(w)$ est le terme d'entropie croisée et E^k = .

Le hessien A de la fonction de coût régularisée se calcule grâce à la formule :

$$A = \quad I(w) =$$

H est la matrice du hessien de la fonction d'entropie croisée dont le calcul exact est effectué grâce à l'algorithme développé par [Bishop, 1992]. Si p est le nombre de paramètres du réseau, le calcul de la matrice H de dimension (p, p) nécessite un nombre d'étapes qui varie comme p^2.

Le calcul du paramètre γ_k peut se faire, soit en inversant la matrice A par une méthode proposée dans [Press *et al.*, 1992], soit en la diagonalisant par une méthode qui permet de calculer les valeurs propres et les vecteurs propres.

La mise en œuvre de cette méthode nécessite de fixer des valeurs initiales pour les hyperparamètres, puis de commencer la minimisation de la fonction de coût. Après un certain nombre d'itérations, la fonction de coût est proche d'un minimum et les hyperparamètres peuvent être calculés à nouveau.

Cependant le minimum trouvé est un minimum de la fonction de coût régularisée et n'est donc pas nécessairement un minimum de la matrice H. La matrice H n'est donc pas nécessairement définie positive au point où l'on calcule son hessien et ses valeurs propres peuvent être négatives. Or, si les valeurs propres sont négatives, il est possible d'obtenir une valeur γ_k négative et donc une valeur négative pour l'hyperparamètre !

Pour remédier à ce problème, il est recommandé dans la FAQ sur les réseaux bayesiens[2] de ne pas prendre, dans la détermination de γ_k, la contribution de ces valeurs.

Lorsqu'on utilise la formule faisant intervenir l'inverse du hessien, il est nécessaire de vérifier les inégalités :

$$0 < \gamma_k \leq p_k$$

Après avoir choisi des valeurs initiales, il faut minimiser partiellement la fonction de coût régularisée, puis estimer les valeurs des hyperparamètres grâce à l'une des formules. L'estimation des hyperparamètres ne doit pas se faire avant une minimisation conséquente de la fonction de coût régularisée, car les approximations n'ont pas de sens loin du minimum. Après avoir modifié les valeurs des hyperparamètres, la surface de la fonction de coût a été modifiée et il est nécessaire de recommencer une nouvelle minimisation partielle.

L'apprentissage se termine après convergence de cet algorithme. En théorie, la convergence de cet algorithme n'a pas été montrée : il est nécessaire de la vérifier expérimentalement.

[2] http://wol.ra.phy.cam.ac.uk/mackay/Bayes_FAQ.html

Dans notre cas, il est nécessaire de ne pas faire varier l'hyperparamètre lié aux biais des neurones cachés, car il diverge systématiquement. Dans ce cas, les biais associés tendent vers zéro très rapidement, mais le produit $\alpha_0 E_w^0[w]$ diverge.

La Figure 8.10 montre l'évolution des deux hyperparamètres α_1 et α_2 pendant l'apprentissage pour les trois thèmes *dlr*, *nat-gas* et *ipi*, chaque point d'une courbe correspond à un nouveau calcul des hyperparamètres. Pour ces trois thèmes, l'hyperparamètre α_2 converge rapidement vers une valeur proche de 0, mais l'hyperparamètre α_1 a tendance à diverger.

Avant chaque calcul, on calcule le conditionnement de la matrice hessienne *A* grâce à une décomposition en valeurs singulières [Press *et al.*, 1992]. L'évolution du logarithme de cette valeur au fil de l'apprentissage est tracée à la Figure 8.11 pour chacun des trois thèmes. Ces courbes montrent que la matrice hessienne est de plus en plus mal conditionnée au fur et à mesure que les hyperparamètres sont calculés, et que, par conséquent, le calcul de l'inverse de cette matrice est de plus en plus instable numériquement : les calculs ne peuvent plus être menés à bien.

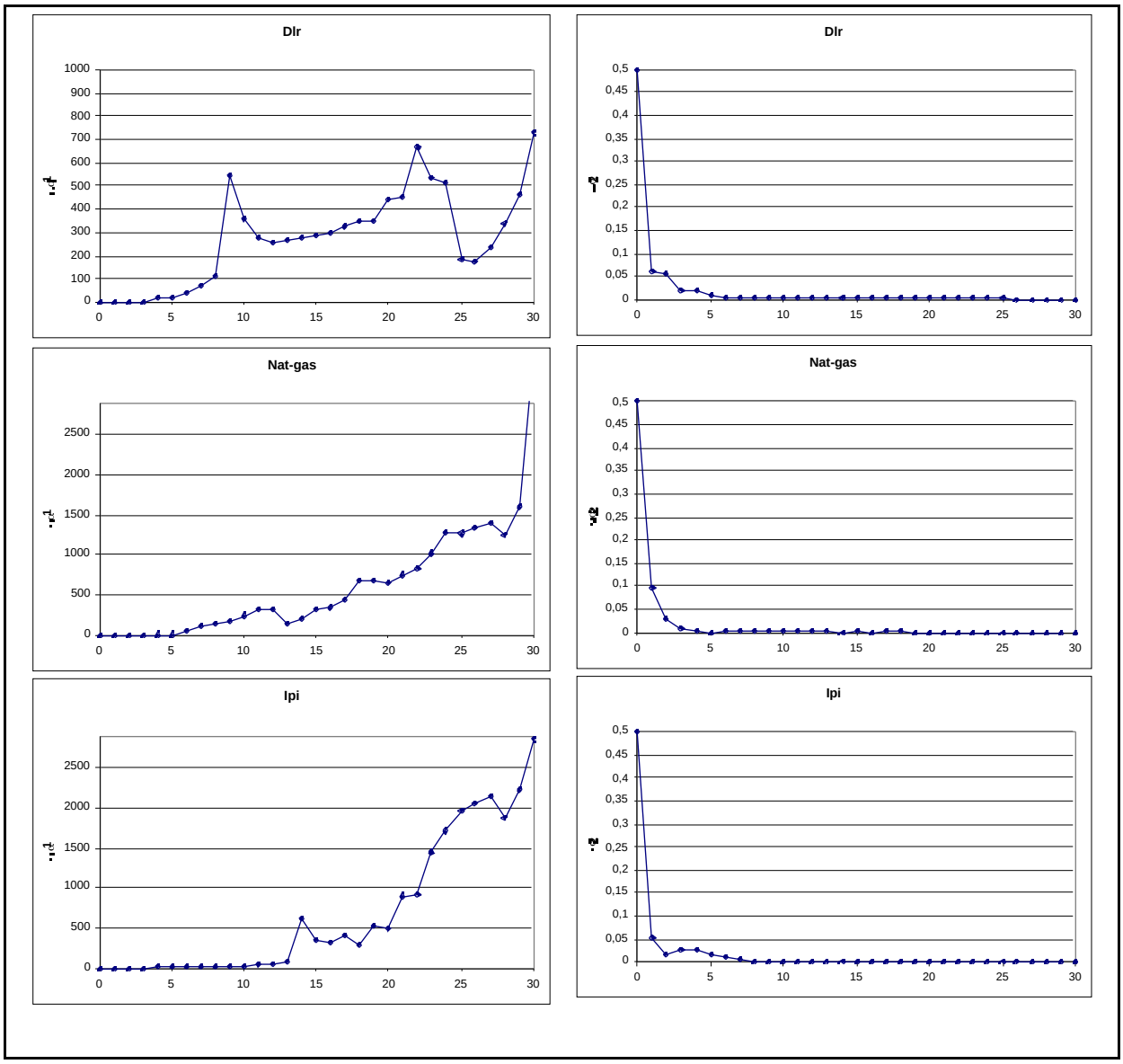

Figure 8.10 : *Évolution des hyperparamètres pendant l'apprentissage.*

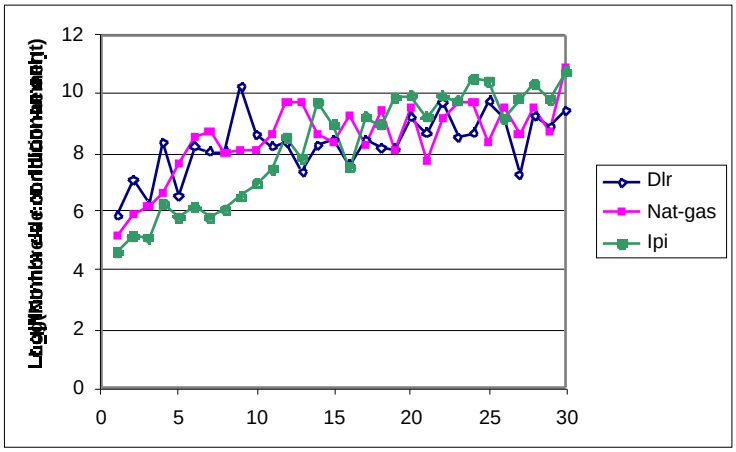

Figure 8.11 : *Évolution du logarithme du nombre de conditionnement de la matrice hessienne A au fil de l'apprentissage.*

Finalement, la méthode de maximisation n'a pas pu être mise en œuvre avec nos modèles, car l'un des hyperparamètres diverge, et rend la matrice du hessien mal conditionnée.
Les formules de calculs des hyperparamètres résultent de plusieurs approximations dont la principale est l'approximation gaussienne de la probabilité *a posteriori* des poids. On peut donc supposer, dans ce cas, que cette approximation n'est pas justifiée.

8.4.4 Conclusion sur les hyperparamètres

Les différentes expériences menées sur le corpus Reuters montrent que les méthodes d'intégration et de maximisation issues de l'approche bayesienne ne conduisent pas à des valeurs correctes pour les hyperparamètres. Il est probable que les approximations nécessaires aux calculs ne soit pas justifiées.

Cependant les résultats du paragraphe 8.4.1 montrent qu'il est possible de choisir *a priori* des valeurs constantes, puisque, d'une part, les résultats ne sont pas, en moyenne, très sensibles à ces valeurs, et, d'autre part, il est possible de retenir des valeurs correctes pour l'ensemble des thèmes. Dans la suite, les paramètres suivants sont utilisés :

$$\alpha_0 = 0{,}001$$

$$\alpha_1 = 1{,}0$$

$$\alpha_2 = 0{,}5$$

Figure 8.12 : *Valeurs des hyperparamètres retenues.*

Il est toujours possible, lorsque les données sont suffisamment nombreuses, de faire de la validation croisée en faisant varier, par exemple, le couple (α_1, α_2). Néanmoins, ces méthodes sont longues à mettre en œuvre, car il est nécessaire d'effectuer un grand nombre d'apprentissages.

8.5 Résultats sur le corpus Reuters

8.5.1 Présentation des expériences réalisées

Plusieurs expériences sont effectuées sur l'ensemble du corpus Reuters afin de mettre en œuvre le modèle proposé, et mesurer l'amélioration apportée par rapport au séparateur linéaire du chapitre 7.

Pour toutes ces expériences, les mots principaux définis à la Figure 8.5 sont les descripteurs sélectionnés lors de la définition du séparateur linéaire du chapitre 7.

Une première série d'expériences est effectuée où, pour chacun de ces descripteurs, les cinq premiers contextes positifs déterminés par la méthode exposée au paragraphe 8.2.1 sont pris en considération.

Pour mesurer l'apport éventuel de l'ajout du contexte négatif, une deuxième série d'expériences est effectuée, où le contexte est défini par les cinq premiers contextes positifs et les cinq premiers contextes négatifs, à condition qu'ils apparaissent au moins dix fois sur la base d'apprentissage.

Par exemple, pour le thème *interest,* le descripteur *rate* avait été sélectionné par la méthode de sélection de descripteurs. Lorsque les cinq premiers contextes positifs sont pris en considération, il peut être désambiguïsé par l'ensemble de descripteurs (*prime*, *cut, base, effective, lending*). Lorsque les contextes négatifs sont également pris en considération, il peut être désambiguïsé par l'ensemble de descripteurs (*prime, cut, base*, *effective, lending, floating, fixed, dollar, rate, exchange*). L'architecture de base de la Figure 8.5 comprend, avec le biais, sept paramètres dans le premier cas et douze dans le second.

Les expériences du chapitre 7 ont montré que la substitution des mots par leur racine lexicale ou par leur lemme augmentait l'ambiguïté des descripteurs et finalement avait un impact négatif sur le modèle linéaire. De nouvelles expériences sont conduites afin de voir si l'introduction du contexte permet de pallier cet inconvénient.

Tous les apprentissages sont effectués avec les valeurs des hyperparamètres indiquées au paragraphe 8.4.4.

8.5.2 Performances du modèle avec contexte

Les performances obtenues sur les 90 thèmes du corpus Reuters sont présentées à la Figure 8.13 ; la première ligne correspond à l'utilisation du contexte positif et la deuxième ligne à l'utilisation des contextes positif et négatif. Les performances sont mesurées grâce à la précision moyenne non interpolée (UAP), la précision moyenne sur 11 points et la mesure de *F* optimisée.

Les résultats sont présentés par sous-ensemble de thèmes pour faire apparaître des corrélations éventuelles entre le nombre de documents pertinents de la base d'apprentissage ; les résultats sont des macro-moyennes sur ces sous-ensembles.

	Nombre de poids du réseau	UAP	*F*	11-pt
Moyenne sur l'ensemble des thèmes	124,76 136,69	72,28 72,89	65,61 66,07	72,44 73,12
Moyenne pour les thèmes 1 à 10	202,60 223,70	90,30 90,29	86,27 86,48	87,16 87,12
Moyenne pour les thèmes 11 à 40	136,16 151,81	84,73 85,01	83,32 83,23	85,18 85,39
Moyenne pour les thèmes 41 à 60	139,95 150,05	70,45 72,26	68,00 70,07	71,48 73,46
Moyenne pour les thèmes 61 à 90	77,93 84,53	55,52 55,85	39,76 39,76	55,87 56,39

Figure 8.13 : *Ensemble des résultats sur le corpus Reuters. La première ligne correspond à l'utilisation du contexte positif et la deuxième ligne à l'utilisation des deux contextes.*

La prise en compte du contexte négatif n'améliore pas les performances alors qu'elle nécessite plus de paramètres.

La seule différence significative dans les moyennes se trouve pour les thèmes 41 à 60 et s'explique en grande partie par le thème *retail* (thème 41) pour lequel la précision moyenne non interpolée évolue de 35,0 à 62,5 grâce à l'ajout du contexte négatif. Cependant, ce thème ne contient que deux documents pertinents sur la base de test : les mesures sont donc très sensibles au moindre changement. La Figure 8.14 montre le classement des dix premiers textes de la base de test obtenu avec chacune des méthodes ; les deux textes pertinents apparaissent

en gras. Le classement des deux méthodes est quasiment identique et il n'est pas possible de tirer des conclusions à partir de cet exemple.

15546	0.777016	**15742**	**0.656229**
15742	**0.636757**	15546	0.589069
17723	0.361686	17723	0.330273
15619	0.298192	15619	0.323393
16853	0.0962382	16853	0.0880126
19625	0.0550316	19625	0.0616612
15033	0.0374317	16783	0.0400094
19483	0.0303551	**16118**	**0.0383093**
16843	0.0292854	15033	0.0380672
16118	**0.0286407**	16843	0.02986

Figure 8.14 : *Liste des dix premiers textes classés pour le thème retail. Colonne de gauche avec le contexte positif, colonne de droite avec les deux contextes.*

La Figure 8.15 présente, pour les soixante premiers thèmes, les comparaisons thème à thème : chaque point représente un thème dont l'abscisse est la précision moyenne non interpolée obtenue avec l'utilisation du contexte positif et l'ordonnée est celle obtenue avec l'utilisation des contextes positif et négatif.

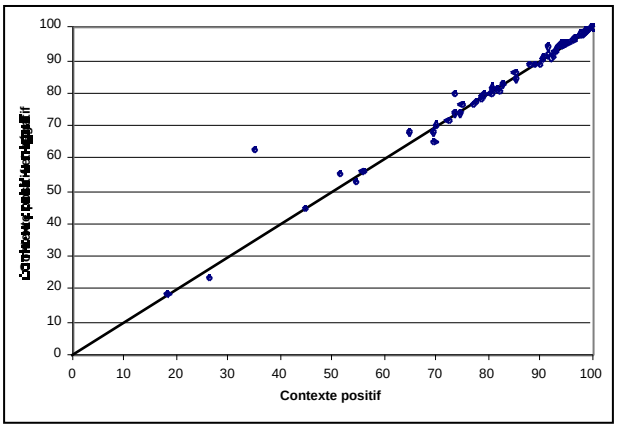

Figure 8.15 : *Comparaison thème à thème pour les soixante premiers de la précision moyenne non interpolée : contexte positif ou contexte positif et négatif.*

Le seul point éloigné de la diagonale correspond au thème *retail* étudié plus haut ; pour tous les autres thèmes, les points sont très groupés autour de la diagonale : pour la majorité des thèmes, les performances sont quasiment inchangées.

En conclusion, l'ajout du contexte négatif augmente le nombre de poids du réseau de neurones, mais l'amélioration des résultats n'est pas significative.

8.5.3 Comparaison avec le séparateur linéaire

Les résultats obtenus avec l'utilisation du contexte positif sont comparés avec les résultats obtenus par le séparateur linéaire présenté chapitre 7. La Figure 8.16 reprend les résultats déjà obtenus par chacune des approches : la première ligne correspond au modèle avec cinq contextes positifs, et la deuxième ligne correspond au séparateur linéaire du chapitre 7.

	UAP	*F*	11-pt
Moyenne sur l'ensemble des thèmes	72,28 72,59	65,61 66,45	72,44 72,83
Moyenne pour les thèmes 1 à 10	90,30 89,68	86,27 85,35	87,16 87,21
Moyenne pour les thèmes 11 à 40	84,73 83,19	83,32 81,37	85,18 83,58
Moyenne pour les thèmes 41 à 60	70,45 69,20	68,00 69,15	71,48 69,99
Moyenne pour les thèmes 61 à 90	55,52 59,03	39,76 43,79	55,87 59,64

Figure 8.16 : *Résultats sur l'ensemble du corpus Reuters. La première ligne correspond au modèle avec contexte, la deuxième ligne au séparateur linéaire.*

Sur l'ensemble des thèmes, les résultats sont très proches, mais globalement, sur les thèmes comprenant plus de dix documents pertinents sur la base d'apprentissage (les soixante premiers), le modèle avec contexte conduit à des résultats supérieurs, tandis que sur les trente derniers, le modèle linéaire est plus performant.

Cependant, les résultats précédents sont des moyennes et peuvent cacher des différences ; pour préciser les résultats, la Figure 8.17 présente, pour les soixante premiers thèmes, la précision moyenne non interpolée obtenue avec chaque méthode.

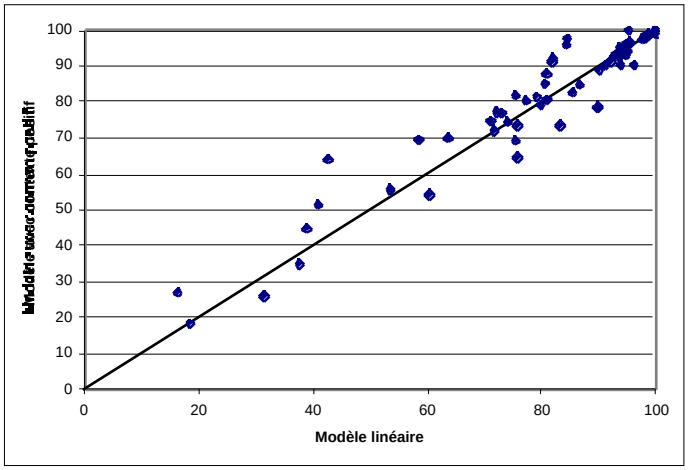

Figure 8.17 : *Comparaison de la précision moyenne non interpolée pour les thèmes 1 à 60 : modèle linéaire ou utilisation du contexte.*

Il est intéressant de noter que, en général, le modèle avec contexte améliore le score des thèmes qui ont une performance plus faible que la moyenne avec le modèle linéaire. Pour les thèmes qui atteignent une précision moyenne non interpolée supérieure à 90 avec le modèle linéaire, l'utilisation du contexte n'apporte pas d'amélioration.

La Figure 8.18 présente, pour chaque modèle, les courbes rappel-précision interpolée pour les soixante premiers thèmes et pour les trente derniers.

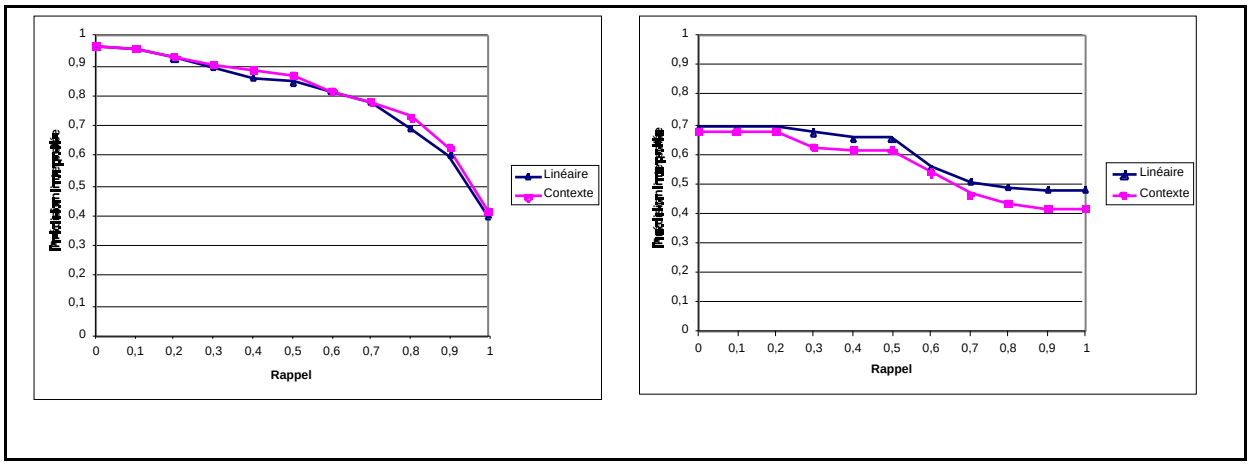

Figure 8.18 : *Courbes rappel-précision interpolée, pour les soixante premiers thèmes (à gauche), et pour les trente derniers (à droite).*

La Figure 8.19 présente les courbes rappel-précision interpolée, pour les trente-six thèmes parmi les soixante premiers pour lesquels la précision moyenne non interpolée est inférieure à

90. Pour l'ensemble de ces thèmes, l'amélioration apportée par l'utilisation du contexte apparaît plus nettement.

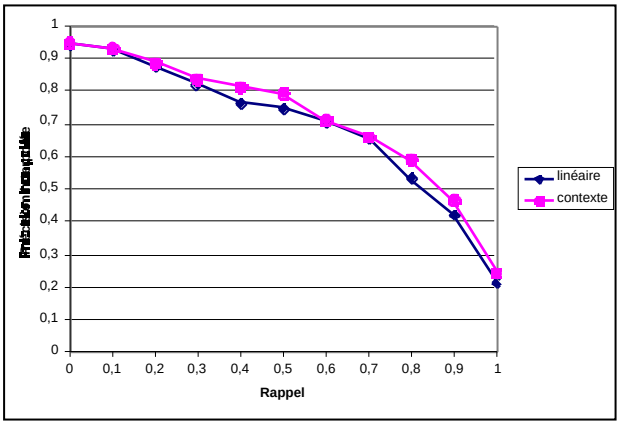

Figure 8.19 : *Courbes rappel-précision interpolée pour les trente-six thèmes dont la précision moyenne non interpolée est inférieure à 90 avec le modèle linéaire.*

8.5.3.1 Exemples de thèmes dont les performances augmentent

Ce paragraphe a pour but d'illustrer les différences de comportements entre les deux méthodes grâce à l'étude de quelques exemples. Dans la discussion, la valeur de la sortie d'un classifieur pour un texte donné est appelée *score*.

Pour le thème *interest* qui contient 347 documents pertinents sur la base d'apprentissage et 131 sur la base de test, les performances obtenues par chacune des méthodes sont présentées à la Figure 8.20 :

	UAP	*F*	11-pt
Modèle linéaire	75,33	71,49	73,49
Modèle avec contexte	81,98	76,68	79,74

Figure 8.20 : *Performance pour le thème interest en fonction du modèle.*

L'observation des sorties de chaque modèle montre qu'avec l'utilisation du contexte, les scores des textes pertinents ont tendance à être plus élevés grâce à l'utilisation d'expression comme *interest rates* qui sont reconnus par le modèle avec contexte. Cette tendance se retrouve sur les courbes rappel-précision de la Figure 8.21, qui montrent que la précision est plus élevée avec le modèle utilisant le contexte.

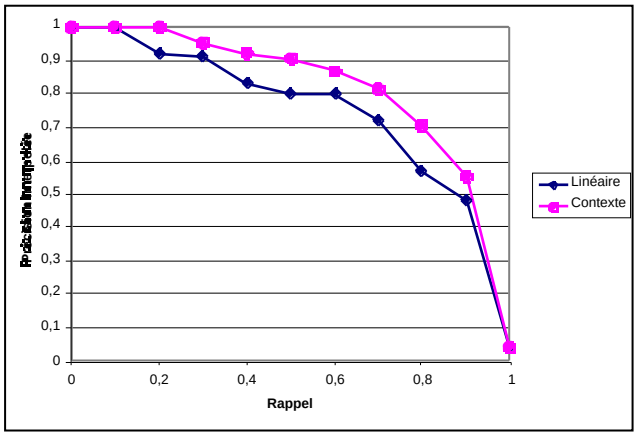

Figure 8.21 : *Courbes rappel-précision interpolée du thème interest pour chacun des modèles.*

Par exemple, le score du texte de la Figure 8.22 passe de 0,40 avec le modèle linéaire à 0,60 avec le modèle utilisant le contexte ; dans ce deuxième cas, aucun document non pertinent n'est classé avant ce texte.

```
AQUINO SAYS MANILA WATCHING INTEREST RATES CLOSELY
President Corazon Aquino said the
Philippines was closely monitoring interest rates in the wake
of Monday's record drop on Wall Street and steep declines in
Manila and other Asian stock markets.
(…)
REUTER
```

Figure 8.22 : *Extrait du texte 20021 pertinent pour le thème interest.*

Pour le thème *cpi* (thème 23, *Consumer Price Index*) qui contient 69 documents pertinents sur la base d'apprentissage et 28 sur la base de test, les performances avec chacune des méthodes sont présentées à la Figure 8.23 et montrent la supériorité du modèle avec contexte :

	UAP	*F*	11-pt
Modèle linéaire	58,38	53,73	58,75
Modèle avec contexte	69,55	61,54	68,17

Figure 8.23 : *Performance pour le thème cpi en fonction du modèle.*

Le texte 16740 (Figure 8.24), qui est pertinent pour le thème est à la 16ème place avec le modèle linéaire (son score est 0,48) et se trouve au 6ème rang avec le modèle qui utilise le contexte (son score est 0,87). Pour le modèle avec contexte, le mot-clef *consumer* est précisé par *index* lui-même précisé par le mot *price* et *inflation* est précisé par *rate* et *pct.*

NEW ZEALAND CPI RISES 2.3 PCT IN MARCH QUARTER
New Zealand's **consumer** price **index**,
CPI, which measures the rate of **inflation**, rose 2.3 pct in the
quarter ended March 31 against an 8.9 pct rise in the December
1986 quarter and a 2.3 pct rise in the March 1986 quarter, the
Statistics Department said.
(…)
Nearly half the increase in the latest quarterly **index** was
contributed by the housing group, the department said.

Figure 8.24 : *Extrait du texte 16740 pertinent pour le thème cpi. Les mots en gras sont les mots sélectionnés par la méthode de sélection de descripteurs.*

Le texte 19081 (Figure 8.25) est un nouvel exemple de texte dont le classement est amélioré par le modèle avec contexte, car *consumer* est précisé, ici, par la présence de (*index, pct*), *index* est précisé par (*price, rose*).

```
CANADA CONSUMER PRICE INDEX UP 0.6 PCT IN MAY
Canada's consumer price index rose 0.6
pct in May to 137.8, base 1981, following a 0.4 pct rise in
April and a 0.5 pct rise in May 1986, Statistics Canada said.
    The May year-on-year rise was 4.6 pct, compared with a 4.5
pct rise in April.
 Reuter
```

Figure 8.25 : *Texte 19081 pertinent pour le thème cpi.*

En revanche, le texte 18001 (Figure 8.26) est un exemple de texte dont le classement diminue puisque son score passe de 0,86 avec le modèle linéaire à 0,52 avec le modèle utilisant le contexte. Pour ce texte, le seul mot-clef présent est *inflation*, mais il n'est accompagné d'aucun contexte pertinent à chaque fois sauf dans le dernier cas avec la présence de *pct.*

```
BRAZIL'S SARNEY RENEWS CALL FOR WAR ON INFLATION
President Jose Sarney today declared "a war without quarter" on inflation
and said the government would watch every cent of public expenditure.
    Sarney, addressing his cabinet live on television, also reiterated that
he intended to remain in power for five years, until 1990. There has been a
long-running political debate about how long his mandate should be.
    Brazil is currently suffering from the worst inflation of its history.
In April monthly inflation reached 21 pct.
 Reuter
```

Figure 8.26 : *Texte 18001 pertinent pour le thème cpi.*

Ce dernier exemple montre que si la performance du thème *cpi* a augmenté grâce à l'utilisation du contexte, certains textes pertinents sont tout de même moins bien classé par l'utilisation du contexte que par le modèle linéaire.

8.5.3.2 Exemples de thèmes dont les performances se dégradent

Pour le thème *hog* (thème 51) qui contient seize documents pertinents sur la base d'apprentissage et six sur la base de test, les performances avec chacune des méthodes sont présentées à la Figure 8.27 :

	UAP	*F*	11-pt
Modèle linéaire	89,72	80,00	82,73
Modèle avec contexte	78,57	71,43	80,09

Figure 8.27 : *Performance pour le thème hog en fonction du modèle.*

Les scores des documents obtenus avec chaque modèle sont présentés à la Figure 8.28. Pour les trois premiers textes pertinents, le modèle avec contexte renforce leur score grâce à l'utilisation de la présence du mot *slaughter* dans le contexte de *hog* comme le montre l'exemple de texte à la Figure 8.29. Cependant, le score des trois derniers textes diminue, car si le mot *hog* est présent dans ces textes, il n'est jamais entouré d'un contexte pouvant renforcer ce mot : le modèle n'est plus assez sensible.

Modèle linéaire		Modèle avec contexte	
16255	0.886748	16255	0.917718
15532	0.886748	15532	0.917718
17823	0.643343	17823	0.850644
17827	0.363257	17827	0.134609
21367	0.332782	21367	0.131715
19555	0.272811	19555	0.0691451

Figure 8.28 : *Ensemble des scores des textes pertinents pour le thème hog.*

```
HOG AND CATTLE SLAUGHTER GUESSTIMATES
Chicago Mercantile Exchange floor
traders and commission house representatives are guesstimating today's hog
slaughter at about 280,000 to 300,000 head versus
294,000 week ago and 303,000 a year ago.
    Cattle slaughter is guesstimated at about 120,000 to 126,000 head
versus 120,000 week ago and 124,000 a year ago.
 Reuter
```

Figure 8.29 : *Texte 16255 pertinent pour le thème hog.*

Cet exemple illustre bien l'un des problèmes qui se pose quand le contexte est pris en considération : certains mots voient leur influence diminuer s'ils ne sont pas entourés d'un contexte pour les désambiguïser. Si ce cas se présente trop souvent, les performances du modèle avec contexte deviennent inférieures à celles du séparateur linéaire.

8.5.4 Représentation des textes avec les racines lexicales

La représentation des textes avec les racines lexicales avait conduit à une dégradation des résultats avec le séparateur linéaire (cf. chapitre 7), notamment à cause de l'ambiguïté ajoutée par l'utilisation des racines. Or, comme le modèle avec contexte effectue une désambiguïsation des mots, on peut espérer tirer profit de l'utilisation des racines sans en subir les conséquences négatives.

Par exemple, la racine *custom*, qui est un descripteur utilisé dans le thème *interest* (cf. chapitre 6), est maintenant précisé par le contexte (*repurchas, reserv, feder, fed, via*) qui sont les racines respectives de *repurchase*, *reserves*, *federal, fed* et *via* ; dans beaucoup de cas, il ne risque plus d'être confondu avec d'autres significations de *custom*.

8.5.4.1 Résultats sur l'ensemble des thèmes

La Figure 8.30 présente les résultats obtenus sur l'ensemble du corpus Reuters avec l'utilisation de racines. Comme précédemment, la première ligne correspond à l'utilisation du contexte positif et la deuxième ligne à l'utilisation des contextes positif et négatif.

	Nombre de poids du réseau	UAP	*F*	11-pt
Moyenne sur l'ensemble des thèmes	108,38 125,92	74,71 73,80	66,64 65,81	74,93 74,20
Moyenne pour les thèmes 1 à 10	183,40 214,6	89,91 89,76	86,20 86,56	86,76 86,64
Moyenne pour les thèmes 11 à 40	112,03 132,90	83,95 83,88	83,07 82,84	84,30 84,51
Moyenne pour les thèmes 41 à 60	111,55 126,20	70,45 70,40	68,34 69,20	71,76 71,79
Moyenne pour les thèmes 61 à 90	78,70 90,43	63,69 61,09	42,87 39,98	64,16 61,77

Figure 8.30 : *Ensemble des résultats avec les racines lexicales sur le corpus Reuters. La première ligne correspond à l'utilisation du contexte positif et la deuxième ligne à l'utilisation du contexte négatif.*

Comme précédemment, l'utilisation du contexte négatif fait légèrement diminuer les résultats tout en augmentant le nombre de paramètres du modèle.

La comparaison de ces résultats avec les résultats de la Figure 8.13 montre que la performance globale est améliorée avec l'utilisation de racines. Cependant l'analyse par sous-ensemble de thèmes montre que les résultats ne sont améliorés que sur le sous-ensemble des thèmes 61 à 90, c'est-à-dire sur l'ensemble des thèmes comportant moins de dix documents pertinents sur la base d'apprentissage.

La Figure 8.31 représente, pour les soixante premiers thèmes, la précision moyenne non interpolée obtenue en utilisant uniquement le contexte positif sans l'utilisation de racines (représentation "brute") et avec l'utilisation de racines.

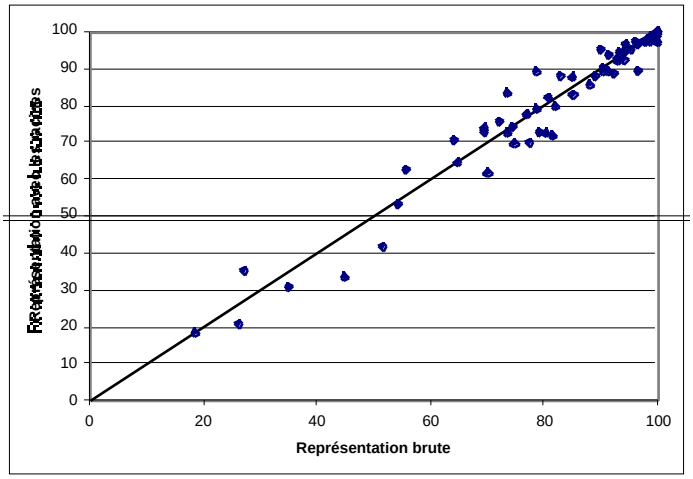

Figure 8.31 : *Comparaison thème à thème de la précision moyenne non interpolée pour les soixante premiers thèmes : représentation brute et avec utilisation de racines.*

Les points sont répartis de chaque côté de la diagonale, mais en moyenne, les performances sont supérieures avec la représentation "brute".

L'utilisation du contexte a permis une désambiguïsation partielle et une amélioration des performances par rapport au séparateur linéaire utilisant les racines, comme le prouvent les comparaisons des performances notamment sur le sous-ensemble des thèmes 41 à 60 (Figure 8.32).

	UAP	*F*	11-pt
Modèle linéaire utilisant les racines	67,5	65,4	68,3
Modèle avec contexte utilisant les racines	70,45	68,34	71,76

Figure 8.32 : *Comparaison des performances sur le sous-ensemble de thèmes 41 à 60 entre le modèle linéaire utilisant les racines et le modèle avec contextes utilisant les racines.*

Cependant, cette désambiguïsation n'est pas suffisante puisque sur les soixante premiers thèmes, les performances sont meilleures lorsque les mots ne sont pas substitués par leur racine.

8.5.4.2 Analyse des thèmes 61 à 90

La variance des résultats est très grande sur le sous-ensemble des thèmes 61 à 90 ; la Figure 8.33 reprend les résultats obtenus sur ce sous-ensemble avec le séparateur linéaire, avec le modèle utilisant les contextes positifs, et avec le modèle utilisant les contextes positifs et les racines. La Figure 8.34 présente les courbes rappel-précison pour les trois modèles.

	UAP	*F*	11-pt
Modèle linéaire	59,03	43,79	59,64
Modèle avec contexte	55,52	39,76	55,88
Modèle avec contexte et racines	63,69	42,87	64,16

Figure 8.33 : *Comparaison des performances sur le sous-ensemble des thèmes 61 à 90.*

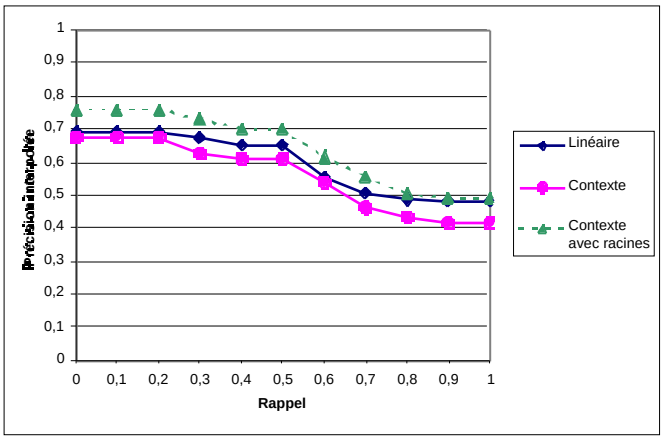

Figure 8.34 : *Courbes rappel-précision interpolée pour le sous-ensemble des thèmes 61 à 90.*

Selon le critère de la précision moyenne non interpolée, l'utilisation du contexte avec les stems est le meilleur des modèles pour ce sous-ensemble de textes. Cependant, selon la mesure *F*, les différences sont moins grandes, car comme on l'a signalé au chapitre 7, le classement des textes peut être parfait, sans qu'il soit possible d'en tirer profit.

Par exemple, la Figure 8.35 présente les résultats obtenus avec chacune des méthodes sur le thème *palmkernel* (thème 78) qui contient deux documents pertinents sur la base d'apprentissage et un document pertinent sur la base de test. Les différences mesurées avec la précision moyenne non interpolée sont très élevées, mais la mesure de *F* montre qu'il est, en fait, impossible de séparer les documents pertinents des autres documents : aucun des modèles n'est satisfaisant.

	UAP	*F*	11-pt
Modèle linéaire	0,03	0,06	0,03
Modèle avec contexte	11,11	0,06	11,11
Modèle avec contexte et racines	100	0,06	100

Figure 8.35 : *Résultats pour le thème palmkernel.*

Finalement, il n'est pas possible de tirer de conclusions sur le bien-fondé de l'utilisation de racines à partir de l'analyse de ce sous-ensemble à cause de la variance des résultats due au faible nombre de documents pertinents sur la base de test. De plus, pour ce sous-ensemble, les performances sont également plus sensibles aux variations de la valeur des hyperparamètres pour les modèles avec contexte : les différences peuvent être dues au choix de ces valeurs plutôt qu'au choix des descripteurs d'entrées.

8.5.5 Représentation des textes avec les lemmes

Le modèle utilisant le contexte a été appliqué comme précédemment, mais en utilisant les lemmes et non plus les racines lexicales comme au paragraphe précédent.

La Figure 8.36 présente les résultats : comme précédemment la première ligne correspond à l'utilisation du contexte positif et la deuxième à l'utilisation des contextes positif et négatif.

	Nombre de poids du réseau	UAP	*F*	11-pt
Moyenne sur l'ensemble des thèmes	112,66 127,91	74,58 74,79	67,08 67,87	74,51 74,89
Moyenne pour les thèmes 1 à 10	187,70 214,3	89,82 90,21	85,52 86,33	86,82 87,43
Moyenne pour les thèmes 11 à 40	115,06 133,77	84,02 84,05	81,73 82,75	84,07 84,25
Moyenne pour les thèmes 41 à 60	124,25 136,55	71,44 73,05	68,86 70,46	72,07 73,99
Moyenne pour les thèmes 61 à 90	78,50 88,66	62,65 62,01	45,50 45,49	62,93 62,41

Figure 8.36 : *Ensemble des résultats avec les lemmes sur le corpus Reuters. La première ligne correspond à l'utilisation du contexte positif et la deuxième ligne à l'utilisation des deux contextes.*

La comparaison des résultats avec les résultats de la Figure 8.30 obtenus avec les racines montre que, si l'on se réfère à la précision moyenne non interpolée, l'utilisation de lemmes est préférable à l'utilisation des racines, mais si l'on se réfère à la mesure de *F*, l'utilisation des racines est préférable.

La comparaison thème à thème des précisions moyennes non interpolées, pour les soixante premiers thèmes, est présentée à la Figure 8.37. Les points sont répartis de chaque côté de la diagonale, et seul deux thèmes présentent des différences de performances majeures.

Pour le thème *retail* (thème 41), la précision moyenne non interpolée passe de 35,0 à 0,06, mais comme on l'a précisé précédemment, ce thème ne comporte que deux documents pertinents sur la base de test : les variations de performances ne sont pas significatives.
Pour le thème *soy-oil* (thème 53) qui comporte quatorze documents pertinents sur la base d'apprentissage et onze sur la base de test, la performance passe de 26,1 sans l'utilisation de lemmes à 68,2 avec l'utilisation de lemmes. L'amélioration, pour ce thème, est réelle, mais c'est le seul thème avec une amélioration si nette.

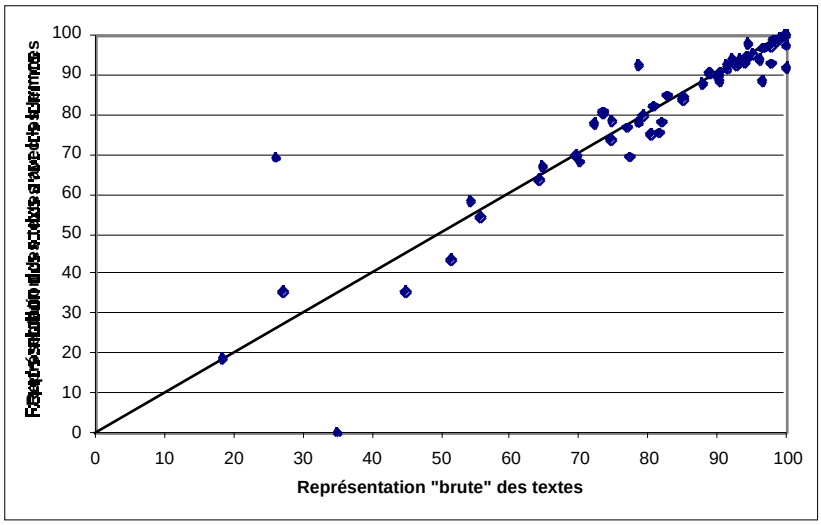

Figure 8.37 : *Comparaison de la précision moyenne non interpolée thème à thème pour les 60 premiers : sans utilisation de lemmes et avec utilisation de lemmes.*

8.5.6 Comparaison des performances obtenues sur Reuters avec d'autres études

Le corpus Reuters a souvent été utilisé comme corpus dans les publications afin de tester différentes approches. Il est possible de comparer les performances obtenues par notre méthode avec les performances obtenues par d'autres approches sur ce même corpus. Ces comparaisons sont intéressantes, car on peut supposer que chaque auteur maîtrise bien la technique qu'il met en œuvre, et qu'il utilise des algorithmes adéquats, ce qui n'est pas toujours le cas lorsqu'une même personne teste des méthodes dont elle n'est pas spécialiste.

Cependant, on ne dispose, pour faire cette comparaison, que de certaines mesures, alors que d'autres critères doivent être pris en considération, comme la marge de progression de chaque méthode, la simplicité de mise en œuvre, ou les temps de calculs.

Il faut noter que la comparaison n'est pas tout à fait objective puisque nous avons largement utilisé la base de test tout au long de ce mémoire pour mieux comprendre le réglage de certains paramètres ; la compétition TREC se prête mieux aux comparaisons objectives.

[Schapire *et al.*, 1998] ont utilisé le corpus Reuters pour tester deux algorithmes : une formule de Rocchio améliorée, et une méthode nommée *AdaBoost* qui est reconnue comme une application efficace de la méthode de *Boosting* [Freund et Schapire, 1997]. Ces deux méthodes sont comparées à l'algorithme *Sleeping Expert* proposé par [Cohen et Singer, 1996]. Les deux méthodes testées dans cet article obtiennent de meilleurs résultats que la méthode *Sleeping Expert*, et sur le corpus Reuters, la formule de Rocchio est meilleure que l'algorithme *Adaboost*, ce qui tend à prouver que son implémentation est très efficace.

Nous avons comparé nos résultats avec ceux obtenus par la méthode de Rocchio. L'implémentation proposée repose sur un codage efficace des fréquences (le codage Lnu) une sélection des documents non pertinents, et une optimisation des poids trouvés. Les comparaisons sont facilitées, car, d'une part, les auteurs considèrent le même découpage du corpus que celui que nous avons utilisé, et, d'autre part, les résultats thème par thème sont disponibles sur le web[3].

[3] http://www.research.att.com/~singhal/sigir98-rocboost.html

La Figure 8.38 présente les résultats de la méthode de Rocchio.

	UAP	*F*
Moyenne sur l'ensemble des thèmes	70,8	56,0
Moyenne pour les thèmes 1 à 10	86,3	80,4
Moyenne pour les thèmes 11 à 40	81,3	73,7
Moyenne pour les thèmes 41 à 60	70,7	59,2
Moyenne pour les thèmes 61 à 90	54,1	27,1

Figure 8.38 : *Résultats de la méthode Rocchio proposé par [Schapire et al., 1998].*

La Figure 8.39 présente les comparaisons thème à thème entre les résultats de la méthode de Rocchio et notre approche (appelée réseaux de neurones) avec l'utilisation du contexte positif : chaque point représente un thème, son abscisse est la précision moyenne non interpolée obtenue par la méthode de Rocchio et son ordonnée est celle obtenue par notre méthode.

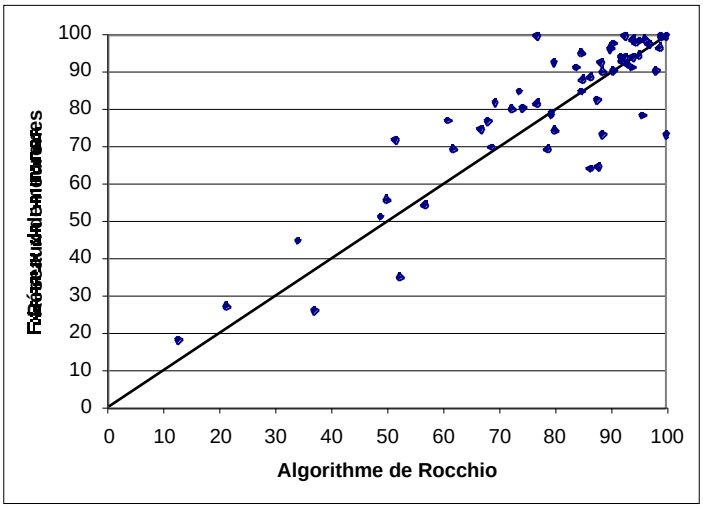

Figure 8.39 : *Comparaison pour les soixante premiers thèmes du corpus Reuters entre la méthode de Rocchio et notre méthode.*

Cette comparaison montre que notre approche est supérieure pour une majorité de thèmes, même si, pour certains thèmes, l'algorithme de Rocchio obtient de meilleurs résultats.

8.5.7 Conclusions des expériences sur le corpus Reuters

Les expériences menées sur le corpus Reuters ont montré que l'utilisation du contexte local améliorait les performances par rapport au modèle linéaire. Il est intéressant de noter que cette amélioration bénéficie surtout aux thèmes dont les performances étaient les plus basses.

L'utilisation du contexte négatif ne diminue pas les résultats, mais elle augmente le nombre de paramètres sans apporter d'amélioration significative.

Malgré la désambiguïsation partielle effectuée grâce à l'utilisation du contexte, la substitution des mots par leur racine ou par leur lemme n'a pas apporté d'amélioration des performances. De plus, bien que ces deux approches réduisent le nombre de paramètres du modèle, elles ajoutent une étape supplémentaire dans la préparation des textes.

8.6 Mise en œuvre sur le corpus TREC-8

Pour tester cette méthode sur TREC-8, on reprend les expériences sur le corpus TREC-8 avec la sélection de descripteurs qui avait abouti au système S2N2.

Pour chacun des descripteurs sélectionnés, on ajoute, dans l'architecture neuronale de la Figure 8.6, les cinq premiers contextes positifs trouvés pour chacun de ces mots.

Les valeurs des hyperparamètres sont choisies *a priori* pour chacun des thèmes :

$$\alpha_0 = 0{,}001$$

$$\alpha_1 = 1{,}0$$

$$\alpha_2 = 0{,}5$$

Les performances sont comparées aux résultats officiels présentés à la compétition (S2N2) et aux améliorations du chapitre 7 (améliorations effectuées après la compétition) apportées à ce modèle appelé *modèle linéaire*.

La Figure 8.40 montre la moyenne des précisions moyennes non interpolées pour chacune de ces méthodes.

	S2N2	Modèle linéaire	Modèle avec contexte
Performance (UAP)	30,7	34,8	38,2

Figure 8.40 : *Précision moyenne non interpolée (UAP) avec ou sans contexte pour l'ensemble des cinquante thèmes du corpus TREC-8.*

Le modèle utilisant le contexte améliore significativement les résultats, puisque l'introduction du contexte a permis une amélioration de 10 % des performances par rapport au modèle linéaire. Par rapport aux résultats de la compétition (S2N2), l'amélioration des résultats est de 24 %.

Les courbes rappel-précision de la Figure 8.41 confirment l'amélioration observée et mettent en évidence la supériorité du modèle utilisant le contexte.

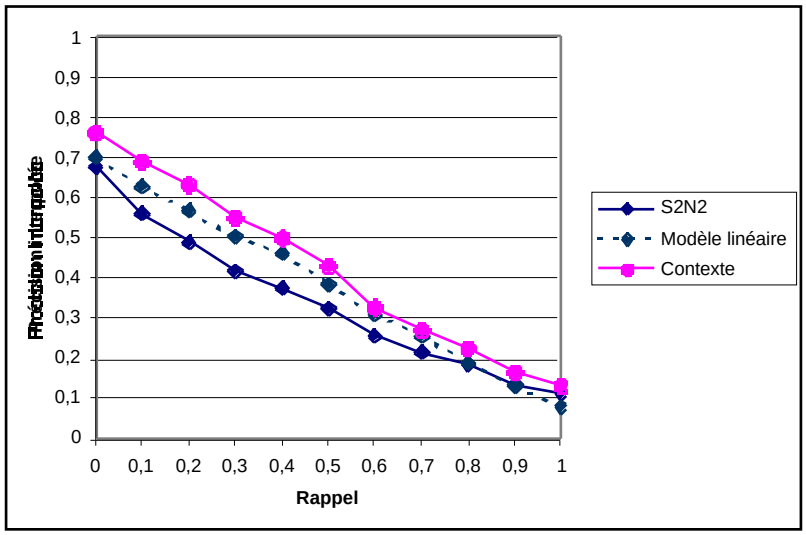

Figure 8.41 : *Comparaison des courbes rappel-précision interpolée pour l'ensemble des thèmes du corpus TREC-8 pour les trois méthodes.*

Enfin, la Figure 8.42 reprend l'ensemble des résultats obtenus par chaque participant pour la compétition ; les nouvelles performances apparaissent sous la dénomination *contexte*. Cette comparaison montre que notre système est maintenant comparable avec les meilleurs systèmes.

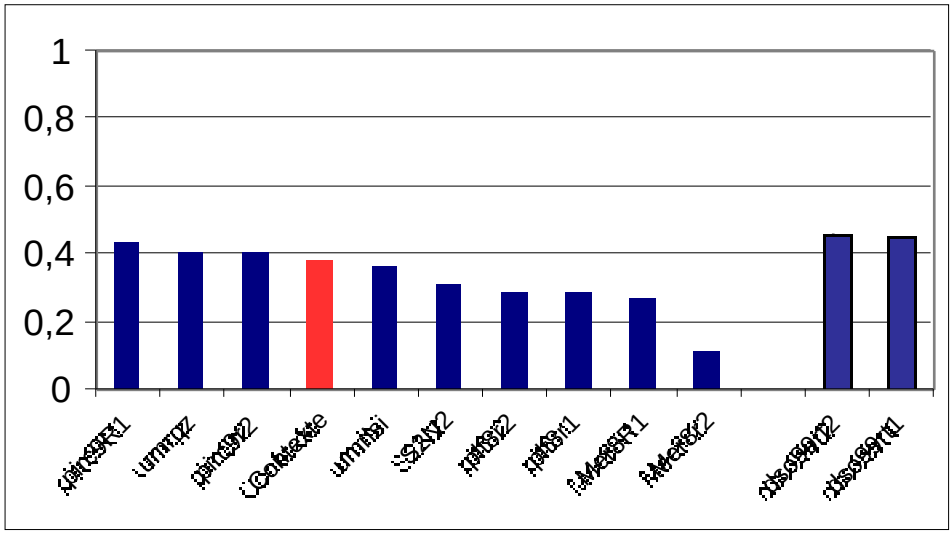

Figure 8.42 : *Résultats de l'ensemble des participants à la tâche de Routing de TREC-8. Les nouvelles performances apparaissent sous la dénomination contexte.*

L'ensemble des résultats confirme les résultats obtenus sur le corpus Reuters : malgré l'augmentation du nombre de paramètres, l'utilisation du contexte améliore les performances.

8.7 Mise en œuvre sur le corpus TREC-9

8.7.1 Détail des paramètres utilisés pour l'apprentissage

Ce paragraphe décrit notre participation à la sous tâche de routing de TREC-9 qui est détaillée dans [Stricker *et al.*, 2001]. La présentation du corpus ainsi que la définition des thèmes est précisée dans l'annexe A.

Pour cette compétition, nous avons mis en œuvre les techniques décrites dans ce chapitre.

Comme le montre la description du corpus à l'annexe A, le nombre de documents pertinents disponibles pour les bases d'apprentissage est beaucoup plus faible sur ce corpus que sur les autres corpus étudiés. Par conséquent, la sélection des descripteurs a été effectuée uniquement par la détermination du vocabulaire spécifique et la méthode d'orthogonalisation de Gram-Schmidt n'a pas été utilisée.

Les vingt-cinq premiers descripteurs trouvés par la méthode du vocabulaire spécifique sont sélectionnés et leurs contextes positif et négatif sont déterminés par la méthode du paragraphe 8.2.1.

Pour chaque thème l'architecture neuronale décrite à la Figure 8.6 contient donc 25 neurones cachés. Le nombre d'entrées liées à chaque neurone caché est défini comme suit : les cinq

premiers contextes positifs sont pris en considération s'ils apparaissent dans plus de deux documents pertinents et les cinq premiers contextes négatifs s'ils apparaissent dans plus de dix documents non pertinents.

L'apprentissage est effectué avec la méthode du *weight decay* ; les valeurs des hyperparamètres sont fixées comme précédemment.

8.7.2 Détail des fichiers envoyés pour la compétition

Nous avons présenté trois fichiers de résultats pour la tâche de routing dont les caractéristiques sont reprises à la Figure 8.43 (les définitions de *OHSUMED queries* et *MeSH sample* sont précisées à l'annexe A ainsi que les précisions sur les annotations manuelles)

	Ensemble des requêtes	Utilisation des annotations manuelles
S2RNr1	*OHSUMED queries* (63 thèmes)	non
S2RNr2	*OHSUMED queries* (63 thèmes)	oui
S2RNsamp	*MeSH Sample* (500 thèmes)	non

Figure 8.43 : *Détail des fichiers résultats soumis pour TREC-9.*

Les résultats obtenus par chaque système ont été présentés lors de la conférence TREC-9 qui s'est déroulée du 13 au 16 novembre 2000.

8.7.3 Résultats obtenus sur les 63 thèmes OHSUMED

Comme lors de la conférence TREC-8, les performances sont évaluées par la précision moyenne non interpolée.

La Figure 8.44 expose les résultats obtenus sur les 63 thèmes OHSUMED et la Figure 8.45 compare ces résultats à ceux des autres candidats.

Parmi tous les résultats proposés, seul notre fichier S2RNr2 a utilisé les annotations manuelles ; la performance obtenue avec ce fichier doit donc être différenciée des autres.

Il est intéressant de noter que la meilleure performance est obtenue par la méthode utilisant les annotations manuelles. Ce résultat prouve que les méthodes à base d'apprentissage numérique peuvent très facilement tirer parti d'informations autres que le texte lui-même sans qu'il soit nécessaire de changer quoique ce soit à la méthode.

Notre fichier S2RNr1 qui n'utilise pas les annotations manuelles est directement comparable à tous les autres participants ; les résultats montrent que notre approche conduit aux meilleures performances.

	S2RNr1	S2RNr2
Moyenne des précisions moyennes non interpolées	0,343	0,385
Nombre de thèmes où notre score est le meilleur	9/63 (14%)	29/63 (46%)
Nombre de thèmes où notre score est supérieur à la médiane	61/63 (97%)	61/63 (97%)

Figure 8.44 : *Résultats pour la tâche de routing de TREC-9 sur les 63 thèmes OHSUMED.*

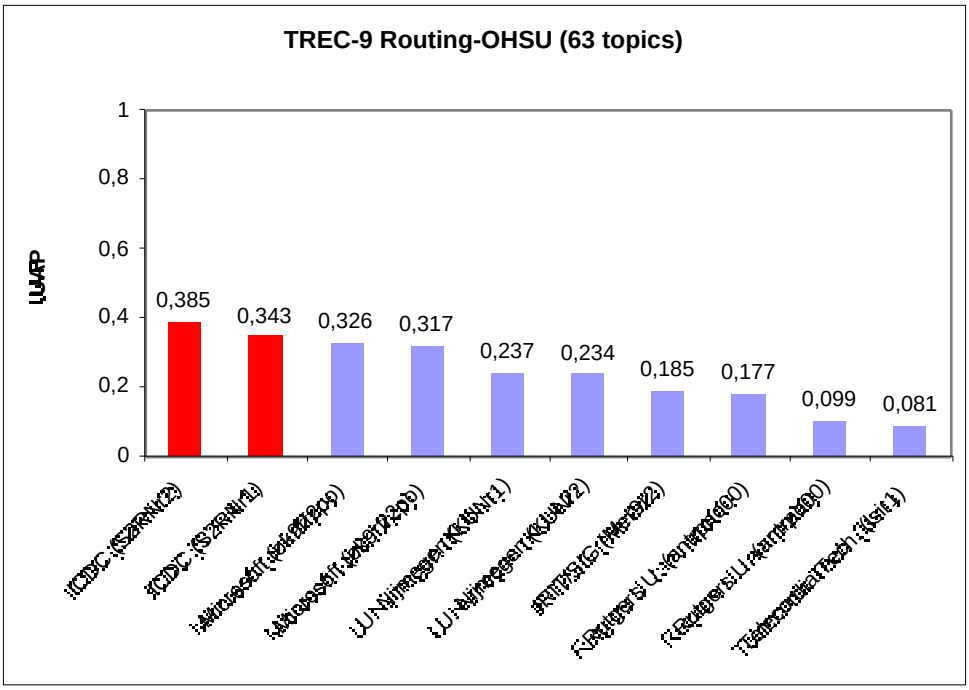

Figure 8.45 : *Moyennes des précisions moyennes non interpolées de l'ensemble des participants à la tâche de routing de TREC-9 pour les 63 thèmes OHSUMED.*

8.7.4 Résultats obtenus sur les 500 thèmes MeSH.

Comme précédemment, les performances sont évaluées par la moyenne des précisions moyennes non interpolées calculées sur chaque thème.

La Figure 8.44 expose les résultats obtenus sur les 500 thèmes MeSH et la Figure 8.45 compare ces résultats à ceux des autres candidats.

	S2RNsamp
Moyenne des précisions moyennes non interpolées	0,335
Nombre de thèmes où notre score est le meilleur	364/500 (73%)
Nombre de thèmes où notre score est supérieur à la médiane	494/500 (99%)

Figure 8.46 : *Résultats pour la tâche de routing de TREC-9 sur les 500 thèmes MeSH sample.*

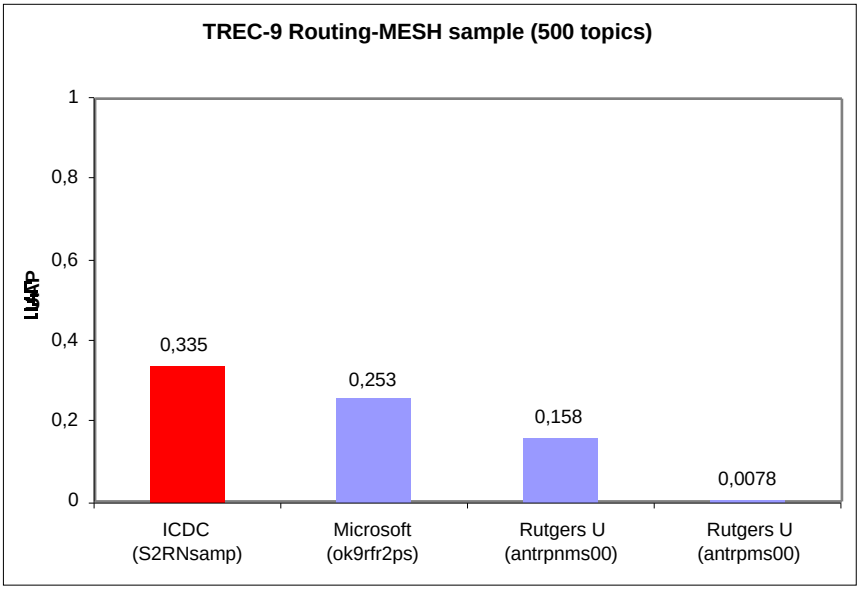

Figure 8.47 : *Moyenne des précisions moyennes non interpolées de l'ensemble des participants à la tâche de routing de TREC-9 pour les 500 thèmes MeSH sample.*

Sur cet ensemble de thèmes, notre approche obtient également les meilleurs résultats. Les écarts sont plus importants que ceux observés précédemment puisque notre système obtient une amélioration de 32 % par rapport au système suivant.

8.7.5 Revue des différents systèmes

Il n'est pas encore possible faire une présentation précise des différentes approches présentées[4] ; nous en présentons juste le principe de base.

Les fichiers **ok9rf2po**, **ok9rfrpo**, et **ok9rfr2ps** ont été présentés par Microsoft et reposent tous sur le modèle OKAPI [Robertson et Sparck Jones, 1976] [Robertson et Walker, 2000] qui est une implémentation du modèle probabiliste présenté au chapitre 2.

Les fichiers **KUNr1** et **KUNr2** ont été présentés par l'université de Nijmegen. Leur approche utilise l'algorithme appelé *Winnow*.

Le fichier **Mer9r2** a été proposé par l'université de Toulouse III. Il s'agit d'une variante du modèle appelé Mercure [Boughanem *et al.*, 2000] qui utilise un réseau de neurones pour implémenter le modèle probabiliste.

Les fichiers **antrpno00**, **antrpo00**, et **antrpnms00** ont été proposés par l'université Rutgers qui a développé un algorithme d'apprentissage appelé *Logical Analysis of Data* (LAD).

Nous n'avons pu trouver aucune information sur le système utilisé pour produire le fichier **lsir1**.

8.7.6 Conclusion

Les résultats obtenus sur la tâche de routing de TREC-9 montrent que notre approche est performante puisqu'elle a été classée première aussi bien sur l'ensemble des 63 thèmes que sur l'ensemble des 500 thèmes.
Sur l'ensemble des 63 thèmes le nombre moyen de documents pertinents disponibles pour la base d'apprentissage est de 10,6 ; donc même avec un nombre aussi faible il est possible de mettre en œuvre notre approche.

[4] Les articles complets seront disponibles sur le site internet de TREC (http://trec.nist.gov) à partir du mois de février 2001.

Les écarts entre notre système et le suivant sont plus élevés sur l'ensemble MeSH (500 thèmes) que OHSUMED (63 thèmes), alors que les deux fichiers ok9rfr2po (0,317 sur OHSUMED) et ok9rfr2ps (0,253 sur MeSH) correspondent au même système avec les mêmes paramètres. Comme sur l'ensemble MeSH le nombre moyen de documents pertinents pour la base d'apprentissage est de 46, cela prouve que notre système a su tirer avantage de ce plus grand nombre de documents pertinents.

Enfin, notre méthode est suffisamment rapide pour pouvoir être utilisée sur un ensemble de 500 thèmes en un temps limité. En effet, la liste des 500 thèmes a été délivrée 14 jours avant la date limite d'envoi des résultats et peu de participants ont fourni des résultats sur cet ensemble.

8.8 Conclusion

Nous avons proposé une extension du séparateur linéaire du chapitre 7, qui consiste à utiliser le contexte local des descripteurs sélectionnés pour effectuer une désambiguïsation. Cette méthode de désambiguïsation repose sur l'étude du corpus étudié pour définir le contexte pertinent d'un mot.

Comme le nombre de paramètres du modèle augmente considérablement par rapport au modèle linéaire, l'utilisation d'une méthode de régularisation est indispensable pendant l'apprentissage. Grâce à l'utilisation du *weight decay*, l'apprentissage est effectué efficacement même pour les thèmes comportant peu d'exemples pertinents.

Les méthodes issues de l'approche bayesienne n'ont pas permis, sur notre problème, de trouver des valeurs adéquates pour les hyperparamètres. Cependant, la méthode la plus simple, qui consiste à fixer les valeurs des hyperparamètres *a priori* donne de bons résultats. Cette méthode qui n'est pas optimale présente le grand avantage de ne pas ajouter de calculs ni de nécessiter de méthodes de validations croisées et donc n'allonge pas la durée de l'apprentissage. Ce paramètre peut être crucial quand il existe beaucoup de thèmes à traiter comme pour TREC-9, le nombre de profils à construire s'élève à cinq cents.

Finalement, les mises en œuvre sur les corpus Reuters, TREC-8 et TREC-9 ont montré que cette approche améliore les performances par rapport au modèle linéaire utilisé précédemment, et que, sur ces trois corpus, les résultats sont comparables aux meilleurs résultats publiés.