ADAPTATION NON SUPERVISÉE DES MODÈLES DE LANGAGE POUR LE SOUS-TITRAGE DE BULLETINS DE NOUVELLES

Jean-François Beaumont

Département de Génie Électrique et Informatique
Université McGill
Montréal, Canada

Avril 2004

Mémoire soumis à la Faculté des Études Graduées et de la Recherche pour répondre en partie aux exigences du grade de Maître Ingénieur.

Your file Votre référence
ISBN: 0-612-98511-3
Our file Notre référence
ISBN: 0-612-98511-3

RÉSUMÉ

Ce mémoire présente une approche permettant de créer des modèles de langage thématiques. Une augmentation de 5% (absolue) du taux de reconnaissance a été obtenue à l'aide de modèles de langage thématiques sur des extraits de nouvelles télédiffusées. Ce document présente d'abord la théorie soutenant les modèles de langage et le problème du manque de données qui existe systématiquement lorsqu'on s'attaque à ce problème. Il est également question des méthodes et des choix qui existent dans la construction et l'adaptation de modèles de langage. Enfin, des expériences sont présentées et commentées afin de démontrer la pertinence du concept de modèles de langage thématiques.

ABSTRACT

This thesis presents an approach to create topic dependent language models. It is shown that a gain of 5% was reached using speech recognition for news broadcast. First, this document presents the theory on which language models are based and the problem of sparse data which is one of the biggest problems associated with the creation of adequate language models. Next, general guidelines are presented in regard of the choices and techniques implied in the creation and the adaptation of language models. Finally, experimental results are presented and commented on to sustain the concept of topic dependent language models.

REMERCIEMENTS

Ce mémoire a été réalisé grâce à la contribution de plusieurs personnes et je tiens à les en remercier. D'abord, Gilles Boulianne du Centre de Recherche Informatique de Montréal (CRIM). C'est par son contact et son esprit scientifique que la simple idée de poursuivre des études graduées s'est matérialisée. Aussi, je suis reconnaissant à toute l'équipe qui forme le groupe de reconnaissance de la parole au CRIM et en particulier à Claude Chapdelaine pour ses bonnes idées et ses encouragements constants. Je ne saurais passer sous silence la collaboration de mon superviseur David A. Lowther du département de génie électrique de l'Université McGill qui m'a permis de travailler sur un sujet de recherche très intéressant. Je tiens également à remercier mon co-superviseur, le docteur Douglas O'Shaughnessy de l'Institut National de Recherche Scientifique (INRS) pour sa patience et son implication dans la rédaction de ce mémoire. Aussi, je tiens à remercier tous les correcteurs, amis et collègues qui ont lu et fourni des commentaires précieux sur ce document. Enfin, c'est grâce au soutien de plusieurs personnes que ce projet est devenu réalité. Je remercie en particulier mes parents, Jacques et Monique, pour leurs constants encouragements au cours de ces dernières années. J'aimerais également remercier Vanessa pour son support et son affection. Ces quelques lignes ne sauraient traduire toute la gratitude et l'amour que je lui porte.

TABLE DES MATIÈRES

LISTE DES FIGURES

LISTE DES TABLES

1. INTRODUCTION

Depuis le début des années 80 [CBC 2003], les téléspectateurs canadiens souffrant d'un problème auditif ont été en mesure de regarder en partie la télévision accompagnée de sous-titres. Cependant, aucune solution n'était alors disponible pour ces personnes pour ce qui était diffusé en direct pendant les émissions de nouvelles par exemple. Vers la fin des années 80, le réseau français de Radio-Canada mettait au point à l'aide de chercheurs un système de sous-titrage basé sur une approche de sténotypie [Descout 1992], laissant de côté la technologie de reconnaissance automatique de la parole de l'époque pour plusieurs raisons :

- la variabilité entre les locuteurs ;
- le traitement des nouveaux mots ;
- la distinction difficile entre plusieurs intervenants lorsque ceux-ci parlent en même temps ;
- le traitement d'un signal parfois bruité ;
- la performance insuffisante des ordinateurs de l'époque pour effectuer de la reconnaissance de la parole à grand vocabulaire en temps réel.

Avec l'avancée de la puissance de calcul et le faible coût des ordinateurs, la reconnaissance de la parole est maintenant une solution envisageable pour répondre au problème de sous-titrage en direct [Brousseau 2003]. Comme la réalisation d'une telle tâche est plus aventureuse que la sténotypie puisqu'il est inévitable que des erreurs se produisent au niveau des modèles probabilistes comme les modèles acoustiques et les modèles de langage, il est nécessaire de maximiser leur performance. Plusieurs paramètres de ce genre de tâche peuvent être contrôlés avec l'utilisation des techniques suivantes :

- L'utilisation d'un perroquet qui répète ce qui est dit par les journalistes
  Cette technique offre des plusieurs avantages ; en plus de permettre d'utiliser un nombre limité de modèles acoustiques dépendants du locuteur, la parole qui sera alors fournie au reconnaisseur est assurée d'être exempte de bruits. Le perroquet aussi appelé locuteur est alors en mesure de répéter ce qui est dit tout en rephra-

sant, tenant compte des limitations du système (des mots du vocabulaire, du débit, du style) et en retirant les hésitations et les reprises.

- L'utilisation de modèles de langage thématiques
  Cette technique consiste à créer des modèles de langage qui sont plus proches des sujets abordés dans un domaine de l'actualité. Ces modèles tirent avantage de la similarité du style journalistique dans un domaine particulier, des expressions couramment utilisées, des noms propres, etc.
- La constante mise à jour des modèles probabilistes du système
  Les modèles acoustiques peuvent être mis à jour au début et à la fin de chaque séance de travail d'un locuteur. Aussi, les modèles de langage peuvent être mis à jour afin d'y ajouter de nouveaux mots et de mettre à jour les probabilités des mots déjà connus des modèles de langage.

Aujourd'hui, l'approche par sténotypie rencontre plusieurs difficultés car ce système n'est plus supporté par son concepteur et que le personnel qui est en mesure de le faire fonctionner se fait de plus en plus rare. En effet, la sténotypie n'est plus enseignée dans nos écoles et la formation d'un individu s'effectue sur plusieurs années ce qui engendre des délais et des coûts importants [Lincoln 2003]. De plus, pour des raisons techniques, de nouveaux mots peuvent être ajoutés à ce système mais la modélisation de ces nouvelles données n'est quant à elle, pas mise à jour. Comme ce système n'a pas été convenablement mis à jour depuis plus de 10 ans, ces utilisateurs sont condamnés à certaines acrobaties et à maintenir une liste grandissante de règles de filtrage pour éviter que certaines erreurs se produisent. Malgré tous ces efforts, certaines erreurs persistent depuis plusieurs années et il est impossible pour les utilisateurs d'y faire quoi que ce soit.

Ce qui différencie l'approche par sténotypie de celle utilisant la reconnaissance de la parole est minime. En fait, la plupart des techniques et des approches qui sont présentées dans ce document pourraient s'appliquer également à l'une ou à l'autre. En faitt, ce qui différencie un système basé sur la reconnaissance de la parole plutôt que sur la sténotypie se trouve à l'entrée. Au lieu d'utiliser la parole, la sténotypie utilise directement une séquence de syllabes, ce qui est analogue au travail d'un dictionnaire et d'un modèle de

langage. Seulement, la reconnaissance de la parole utilise des observations acoustiques pour obtenir les phonèmes (l'unité atomique des syllabes). En somme, la variable supplémentaire imposée par la reconnaissance de la parole se situe au niveau de la performance des modèles acoustiques.

1.1 Sujets abordés et buts du mémoire

Bien que tous les aspects de la reconnaissance automatisée de la parole mérite de l'attention, l'intérêt de ce mémoire vise à décortiquer particulièrement la théorie de l'apprentissage par ordinateurs (Machine Learning) appliquée aux corpus de textes. La reconnaissance de la parole est un vaste domaine de recherche et ce document n'en couvrira pas tous les aspects.

Ce mémoire s'intéresse particulièrement au développement de modèles de langage. Après une révision de l'état de l'art sur les algorithmes d'estimation, on abordera les techniques permettant d'adapter les modèles pour refléter une réalité changeante comme celle du milieu des nouvelles. Certaines techniques d'adaptation de modèles de langage seront présentées dans ce document. Finalement, des expériences portant sur de la reconnaissance à grand vocabulaire sur des bulletins de nouvelles québécois seront réalisées. Ce travail a porté sur les modèles de langage thématiques afin d'augmenter le taux de reconnaissance.

1.2 Contenu des chapitres

Le chapitre 2 présente les concepts de base sur la modélisation statistique du langage incluant un exposé des techniques d'estimation. Le chapitre 3 développe la recherche d'informations appliquée aux modèles de langage. Le contexte de recherche est présenté au chapitre 4. Le chapitre 5 expose quant à lui les expériences, les résultats et les interprétations. Finalement, une discussion sur ces résultats, sur le statut de la technologie et sur la logique concernant les travaux futurs est présentée au chapitre 6.

2. LA MODÉLISATION STATISTIQUE DU LANGAGE

La raison principale qui pousse les chercheurs à utiliser les modèles de langage est que la performance des modèles acoustiques n'est pas suffisante pour égaler la performance d'un humain pour une tâche de reconnaissance à grand vocabulaire. En effet, l'être humain utilise beaucoup d'informations ne relevant pas de l'acoustique comme : la syntaxe, la sémantique, la pragmatique le dialogue et la connaissance de son interlocuteur. Ces sources d'information peuvent être adéquatement représentées avec un modèle de langage bien qu'il ne constitue qu'une approximation du langage [Rosenfeld 2000].

Les modèles de langage ont une grande responsabilité dans le comportement général d'un système de reconnaissance de la parole. Tout cet environnement dont l'humain dispose n'est en effet pas accessible à l'ordinateur et un système informatisé doit par conséquent être soigneusement réalisé par un expert qui fera en sorte que les probabilités contenues dans ce modèle reflètent bien la réalité. Une étude [Lippmann 1997] démontre d'ailleurs l'importance des modèles de langage ainsi que la différence importante de la performance qui existe entre l'humain et l'ordinateur : de 2% de taux d'erreur pour l'être humain contre 17% pour un système informatisé.

Cette approche statistique se retrouve sous plusieurs formes mais le type le plus commun est le modèle *n*-grammes. Avec ce type de modèle, il est possible de faire varier la proximité des mots considérés pertinents. Les approches les plus communes sont le bigramme et le trigramme étant donné la quantité de mots qu'ils nécessitent pour leur construction [Seymore 1996b]. Les modèles de langage produisent des observations statistiques sur des séquences de mots sans tenir compte si le langage ou la tâche modélisé est en français ou en anglais ou si la tâche est de la dictée ou de la transcription de paroles spontanées. Pour ces modèles, les suites de mots sont une suite de symboles, rien de plus [Rosenfeld 2000].

2.1 Théorie

Les techniques de modélisation statistique du langage font parties des clés maîtresses de plusieurs systèmes s'intégrant à des domaines de recherche comme la reconnaissance de la parole [Church 1988], la recherche d'information (IR) [Song 1999], la détection et le suivi de sujets [Walls 1999, Jin 1999], la reconnaissance d'écriture [Marti 2001], la traduction automatisée [Sawaf 2000] et la correction orthographique [Srihari 1993].

Dans le domaine de la reconnaissance de la parole, le modèle de langage permet de mettre les mots en contexte, favorisant une suite de mots plus probable qu'une autre selon la fréquence observée pour cette suite de mots. Par exemple, dans un contexte de bulletins de nouvelles $p(bonsoir) \approx 0.0002$ si en moyenne, 1 mot sur 5000 se trouve à être bonsoir, lors de l'ouverture du bulletin. Cependant, $p(boîte\ canard\ maison) \approx 0$ et $p(canard\ achète\ maison) \approx 0$, étant donné qu'il est improbable de rencontrer de telles suites de mots. Contrairement au domaine de la linguistique où on utilise des grammaires afin de déterminer la validité ou la probabilité d'une séquence de mots, la grammaticalité n'a aucune importance pour les modèles de langage. C'est en fait la fréquence et seulement celle-ci qui définit la probabilité qu'une suite de mots obtiendra.

Le modèle de langage s'inscrit en fin de course dans la composition du réseau de recherche. Schématiquement, on peut représenter un réseau de reconnaissance de la manière suivante :

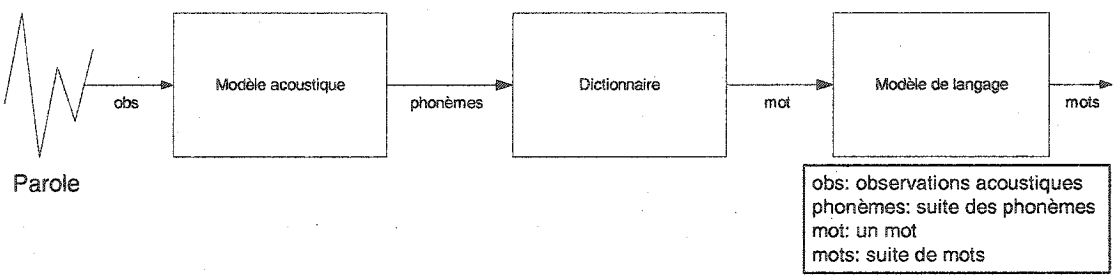

Figure 1: Composantes d'un réseau de reconnaissance

À l'entrée du système se retrouve la parole humaine. Le système établit alors une série d'observations en fonction du temps et c'est le modèle acoustique qui prend en charge la représentation de ces observations en phonèmes. Le dictionnaire quant à lui fait le lien entre une série de phonèmes et un mot faisant partie du vocabulaire du réseau. Le modèle de langage, pour sa part, donne une probabilité pour les suites de mots possibles en favorisant les suites de mots les plus probables. Généralement, ce seront celles qui ont déjà été observées.

Le type de modèle de langage le plus utilisé est le modèle de langage de type *n*-grammes. Pour une séquence *s* composée des mots $w_1,...w_l$, nous pouvons exprimer, sans perte de généralité, $p(s)$ comme :

$$p(s) = p(w_1)p(w_2|w_1)p(w_3|w_1w_2)...p(w_l|w_1...w_{l-1}) = \prod_{i=1}^{l} p(w_i|w_1...w_{i-1}) = \prod_{i=1}^{l} (w_i|h)$$

où *h* représente l'historique, soit les mots précédents w_i dans la phrase. Cet historique fait en sorte que le modèle devient impraticable rapidement car la mémoire des ordinateurs est limitée. Pour pallier à ce problème, on a recours à théorie de Markov afin d'utiliser une longueur fixe d'historique.

Pour fins de démonstration, considérons les unigrammes et les bigrammes. Dans ce cas, $n = 1,2$ respectivement, représentant cet historique. Dans ces cas particuliers, une approximation est faite à savoir que la probabilité d'un mot peut être évaluée en ne considérant d'abord aucun et ensuite seulement que le mot précédent.

$$p(s) = \prod_{i=1}^{l} p(w_i|w_1,...,w_{i-1}) \approx \prod_{i=1}^{l} p(w_i)$$

$$p(s) = \prod_{i=1}^{l} p(w_i|w_1,...,w_{i-1}) \approx \prod_{i=1}^{l} p(w_i|w_{i-1})$$

À partir de *n*=2, il est nécessaire d'introduire des symboles supplémentaires pour donner un sens à $p(w_i|w_{i-1})$ pour $i = 1$. On ajoute donc <s> chargé de représenter w_0. De la même manière, un symbole </s> est ajouté pour marquer la fin d'une phrase pour permettre aux probabilités de toutes les phrases contenues dans le corpus de sommer à 1.

Pour être en mesure de continuer cette introduction à l'aide d'un exemple, prenons le corpus de phrases suivant :

<s> bonsoir mesdames et messieurs </s>
<s> ce soir en manchettes </s>
<s> et maintenant les nouvelles </s>

Table 1: Petit corpus de textes

Pour calculer la probabilité de l'unigramme $p(w_i)$, nous utilisons la fréquence relative $c(w_i)$ du mot w_i en le divisant par le nombre de mots qui se retrouvent dans le corpus, c'est-à-dire $p(w_i) = \frac{c(w_i)}{\sum_{j=1}^{V} c(w_j)}$ où V est la taille du vocabulaire.

La grille de fréquences correspondante pour les unigrammes est présentée à la table 3. C'est la fréquence relative qui détermine la probabilité de l'unigramme :

Unigramme	$c(w_i)$	$p(w_i)$	
</s>	4	4/24	16.67%
<s>	4	4/24	16.67%
bonsoir	1	1/24	4.17%
ce	1	1/24	4.17%
en	1	1/24	4.17%
et	3	3/24	12.50%
les	2	2/24	8.33%
maintenant	2	2/24	8.33%
manchettes	1	1/24	4.17%
mesdames	1	1/24	4.17%
messieurs	1	1/24	4.17%
nouvelles	2	2/24	8.33%
soir	1	1/24	4.17%
	24		100.00%

Table 2: Calcul de la probabilité: unigrammes en détails

Pour calculer la probabilité de la phrase : <s> bonsoir les nouvelles </s>, nous devons donc multiplier chacune des probabilités des unigrammes contenus dans la phrase :

p(<s> ce soir les nouvelles </s>) =	p(<s>)	p(bonsoir)	p(les)	p(nouvelles)	p(</s>)
	$\frac{4}{24}$	$\frac{1}{24}$	$\frac{2}{24}$	$\frac{2}{24}$	$\frac{4}{24}$

Table 3: Calcul de la probabilité à l'aide d'unigrammes

p(<s> ce soir les nouvelles </s>) ≈ 0.000008

C'est d'une manière semblable que la probabilité d'une phrase est calculée à l'aide des bigrammes. Pour la probabilité du bigramme $p(w_i|w_{i-1})$, c'est le nombre de fois que ce bigramme se retrouve dans le corpus divisé par le nombre de fois qu'on retrouve l'unigramme précédent le mot w_i, c'est-à-dire $p(w_i|w_{i-1}) = \frac{c(w_{i-1}w_i)}{c(w_{i-1})}$.

Pour calculer p(<s> ce soir les nouvelles </s>), on doit décomposer la phrase en une succession de paires de mots de la manière suivante :

P(ce|<s>)	$\frac{1}{24}$
P(soir|ce)	$\frac{1}{24}$
P(les|soir)	$\frac{1}{24}$
P(nouvelles|les)	$\frac{1}{24}$
P(</s>|nouvelles)	$\frac{1}{24}$

Table 4: Calcul du maximum de vraisemblance à l'aide de bigrammes

Nous obtenons p(<s> ce soir les nouvelles </s>) ≈0.000013

De la même manière, la grille d'analyse des bigrammes du corpus de texte :

Bigramme	$c(w_i w_{i-1})$	$p(w_i w_{i-1})$	
<s> bonsoir	1	1/4	25.00%
<s> ce	1	1/4	25.00%
<s> et	2	2/4	50.00%
bonsoir mesdames	1	1/1	100.00%
ce soir	1	1/1	100.00%
en manchettes	1	1/1	100.00%
et maintenant	2	2/3	66.67%
et messieurs	1	1/3	33.33%
les nouvelles	2	2/2	100.00%
maintenant les	2	2/2	100.00%
manchettes </s>	1	1/1	100.00%
mesdames et	1	1/1	100.00%
messieurs </s>	1	1/1	100.00%
nouvelles </s>	2	2/2	100.00%
soir en	1	1/1	100.00%
	20		

Table 5: Calcul du maximum de vraisemblance: bigrammes en détails

L'idée à la base de la modélisation du langage est qu'à partir d'un échantillon de textes *m*, on approxime le modèle de langage *M*. Cette approche nécessite un nombre d'exemples impressionnant. En considérant un vocabulaire de grandeur V, le nombre de *n*-grammes qu'un modèle *n*-grammes d'ordre n peut contenir est de V^n. La taille des *n*-grammes couramment utilisée est donc limitée étant donné la mémoire nécessaire et le nombre d'exemples (le nombre de mots) disponibles. En effet, pour un corpus de textes, à mesure qu'on augmente l'ordre d'un modèle de langage, on ne peut observer qu'une fraction de plus en plus petite des *n*-grammes pour lesquels on veut obtenir une probabilité. À cause de ces raisons, les modèles de langage généralement utilisés sont composés d'unigrammes, de bigrammes et de trigrammes ($n = 1,2,3$) [Bourlard 1996]. Enfin, en faisant référence à n, la littérature définit l'ordre du modèle. Cette terminologie vient des modèles de Markov. Fait Enfin, un modèle *n*-gramme peut être interprété comme un modèle de Markov d'ordre $n-1$.

2.1.1 Grammaires et modèles *n*-grammes

En reconnaissance de la parole, une alternative aux modèles de langage *n*-grammes est d'utiliser une grammaire hors contexte ou CFG (Context Free Grammar). Contrairement aux modèles *n*-grammes, la probabilité assignée pour une séquence de mots est dichotomique : soit 1 ou 0. Ces valeurs indiquent si la suite de mots est acceptée ou non selon les spécifications de la grammaire. Ces grammaires sont efficaces dans les contextes où les utilisateurs sont contraints à dire les mots dans un certain ordre. Il existe également des grammaires stochastiques ou SCFG (Stochastic CFG) où une probabilité entre 0 et 1 peut être assignée à une suite de mots en fonction de leur fréquence d'occurrence. Ces grammaires sont particulièrement populaires dans des systèmes d'informations téléphoniques grand public [Knight 2001].

Il est également possible de construire l'équivalent d'un modèle *n*-grammes à partir d'une grammaire [Galescu 1998, Jurafsky 1995]. On peut en effet, générer toutes les possibilités de séquences de mots. De cette façon, on obtient un corpus de textes sur lequel il est possible d'entraîner un modèle *n*-grammes. Ce modèle peut ensuite être adapté à l'aide d'un modèle plus général, dans le but d'opérer une transition entre un style contraint et un style plus libre. En effet, les utilisateurs préfèrent les interfaces vocales construites autour d'un dialogue à initiative mixte, c'est-à-dire où un utilisateur expérimenté peut donner ses instructions rapidement d'un souffle alors qu'un néophyte se laissera diriger de questions en questions.

Les modèles *n*-grammes, malgré leur construction simpliste, puisqu'il n'est pas nécessaire d'analyser un langage et de produire un modèle à la main lui correspondant, s'avère tout de même être une meilleure solution pour un système de reconnaissance de la parole à grand vocabulaire. D'ailleurs, malgré leurs efforts, les linguistes ne sont toujours pas parvenus à écrire une grammaire complète pour l'anglais ou le français [O'Shaughnessy 2003].

2.2 Métrique d'évaluation

Pour évaluer un modèle de langage, les techniques propres à l'apprentissage automatique (Machine Learning) sont utilisées. On sépare le corpus de textes en trois parties distinctes. D'abord le corpus d'entraînement qui permet de créer le modèle de langage, un corpus de développement qui permet de raffiner le modèle et finalement un corpus de test qui permettra de quantifier la qualité du modèle sur les données.

Dans un contexte de reconnaissance de la parole, on s'intéresse évidemment au gain en termes de taux de reconnaissance qu'une modification au modèle de langage apporte. Par contre, comme cette mesure s'avère coûteuse en temps, une alternative intéressante se trouve à être d'évaluer la confusion qui existe dans un modèle de langage sur un corpus de texte. La perplexité vient, en ce sens, quantifier, par le biais de l'entropie, la difficulté qu'aura un système de reconnaissance de la parole à s'acquitter de sa tâche [Jelinek1997].

2.2.1 Perplexité

La perplexité évalue la complexité d'un modèle de langage pour un texte donné. Pratiquement parlant, elle représente l'inverse du facteur de branchement dans le graphe du modèle de langage. Étant donné N, le nombre de mots dans une phrase et $\log_{10} p(s)$ étant le logarithme en base 10 de la probabilité de séquence de mots s, la perplexité se calcule comme suit :

$$ppl = 10^{\frac{\sum_{w_i} \log_{10}(p(s))}{N}}$$

La perplexité est également liée à l'entropie. L'entropie correspond aux nombres de bits nécessaires en moyenne pour encoder chacun des mots du texte d'évaluation. L'entropie d'une séquence de mots s est définie par :

$$H(s) = -\frac{1}{N}\log_2(p(s))$$

La perplexité correspondante est :

$$ppl(s) = 2^{H(s)}$$

La perplexité augmente donc avec l'entropie. La valeur de la perplexité peut varier entre 1 et la taille du vocabulaire. Une perplexité de 1 indique qu'il n'y a qu'un chemin dans le graphe du modèle de langage pour chacun des n-grammes contenus dans une phrase donnée. À l'inverse, une perplexité égale à la taille du vocabulaire indique une grande confusion dans le modèle de langage. En effet, un tel modèle ne favoriserait aucune suite de mots, rendant sa présence inutile.

Un système de reconnaissance de la parole à grand vocabulaire est acceptable avec un modèle de langage ayant une perplexité autour de 100 [Young 1998]. Une réduction de la perplexité de 5% ou moins est rarement significative. Tandis qu'une amélioration (une diminution) de 10% à 20% permet une amélioration intéressante. Cependant il est rare d'obtenir une amélioration de plus de 30% [Rosenfeld 2000].

Malgré l'utilisation répandue de cette métrique pour les modèles de langage, il n'y a pas de correspondance directe entre la perplexité et le gain équivalent du taux de reconnaissance. Il n'est d'ailleurs pas garanti qu'une diminution de la perplexité résultera en une amélioration du taux de reconnaissance [Clarkson 1998].

2.3 Préparation du corpus

Pour obtenir un corpus de textes permettant de construire un modèle de langage, il est par exemple nécessaire d'avoir plusieurs centaines de millions de mots pour évaluer correctement un modèle à base de trigrammes [Rosenfeld 2000]. Idéalement, pour faire de la reconnaissance sur des données spontanées, ces mots proviendraient de transcriptions,

minutieusement réalisées sur une grande quantité de conversations ou d'enregistrements. Dans la pratique cependant, cela n'est pas toujours possible; on ne dispose souvent que de peu de données pour préparer un corpus. En ce sens, il faut alors normaliser le texte disponible afin de limiter la variabilité lexicale et de bien représenter le style de la tâche.

Pour effectuer cette opération de normalisation, quelques méthodes sont proposées ici afin de pallier aux problèmes les plus souvent rencontrés dans une telle tâche : (ces techniques s'appliquent particulièrement aux textes de nouvelles mais l'approche peut s'avérer d'intérêt pour la normalisation de textes en général)

- Identifier les débuts de phrases

  Les débuts et les fins de phrases sont normalement difficilement identifiables. Pour être à même de travailler facilement avec un corpus de textes, il sera profitable de produire un fichier avec une phrase par ligne. Ce formatage peut être difficile lorsque, par exemple, on n'a aucun contrôle sur la source.
- Normaliser le texte en minuscules

  En convertissant tous les mots des textes en minuscules, cela a pour effet de diminuer la taille du vocabulaire. Il faut cependant tenir compte des noms propres car eux ne doivent pas être convertis en minuscules (du moins, pas en entier). Afin de convertir le texte en minuscules, il est nécessaire d'avoir sous la main une liste la plus complète possible de noms propres pour savoir comment corriger le premier mot d'une phrase.
- Les nombres et les dates et heures, les numéros de téléphone

  Comme le corpus de textes est généralement utilisé pour créer le vocabulaire et par la suite le dictionnaire (incluant l'équivalent phonétique des mots), il est important de convertir en texte ces informations afin d'uniformiser la présentation des informations.

- Les abréviations et les unités

  Comme il existe plusieurs façons différentes d'écrire la même abréviation (chacune des lettres séparées par un tiret ou un point), ces abréviations doivent être normalisées afin de limiter la taille du vocabulaire.

Enfin, il ne faut pas perdre de vue que la normalisation du texte permet aux étapes subséquentes de construire des estimateurs valables pour un corpus. Malgré que cette tâche puisse représenter un travail colossal, elle requiert plus ou moins de minutie. Il n'est en effet pas très payant de couvrir les cas moins fréquents. Il est d'ailleurs préférable de s'attarder aux normalisations de séquences de mots les plus fréquentes d'abord. Ce sont elles qui auront le plus d'impact sur la performance de la reconnaissance. Il n'est donc pas nécessaire de normaliser tous les phénomènes du corpus. Cette normalisation doit d'ailleurs avoir pour effet de faire diminuer d'une manière significative la perplexité du modèle de langage. [Adda 1997].

2.3.1 Sélection du vocabulaire

Comme il n'est pas toujours possible d'utiliser tous les mots contenus dans un corpus de textes, il est fréquent de retrouver un mot générique <unk> qui représente à lui seul tous les mots qui ne sont pas contenus dans le vocabulaire. Par exemple, le corpus de La Presse de 2000 à 2003 contient 1.2 millions de mots de vocabulaire. Comme la taille du modèle de langage est proportionnelle à V^n où V est la taille du vocabulaire et n l'ordre du modèle, on doit limiter le vocabulaire afin de contrôler la taille du modèle ainsi créé. Les mots qui ne font pas partis du vocabulaire sont alors des mots considérés hors vocabulaire.

De plus, la sélection du vocabulaire constitue un facteur déterminant dans la mise au point de modèles en reconnaissance de la parole. En effet, chaque mot hors vocabulaire cause en moyenne plus d'une erreur (entre 1.5 et 2 erreurs) de reconnaissance [Pallet 1994]. Il a été montré qu'augmenter un vocabulaire de 20000 mots avec de 1000 à 4000 mots judicieusement choisis peut faire passer la couverture de 93% à 98% [Cole 1996]. Le vocabulaire est généralement déterminé en triant en ordre décroissant les mots par

leurs fréquences pour choisir les plus fréquents comme vocabulaire. Cette technique a pour avantage de rejoindre l'objectif visé par le vocabulaire, c'est-à-dire d'obtenir la meilleure couverture lexicale possible. C'est également une technique qui a pour effet de maximiser la vraisemblance (maximum likelihood). On utilisera alors les données les plus récentes pour faire cette sélection ou des données regroupées par domaine [Gauvain 1995].

2.4 Traitement des événements non vus

Étant donné que le nombre de probabilités à estimer évolue de manière exponentielle en fonction de l'ordre du modèle de langage pour un vocabulaire donné, il n'est pas inédit d'être confronté à des suites de mots non vues. Par exemple, avec le corpus de la Table 1, dans la phrase « les nouvelles ce soir » :

p(les|<s>)	$\frac{0}{4}$

Table 6: Probabilité d'un événement non vu

Ce qui fait que p(<s> les nouvelles ce soir </s>) obtient une probabilité de 0. Pourtant, cette phrase risque de se produire et l'estimateur n'est pas en mesure de fournir une probabilité non nulle.

2.4.1 Méthode de repli (backoff)

La méthode de repli ou backoff [Katz 1987] consiste à utiliser l'ordre inférieur d'un *n*-gramme lorsque celui-ci n'est pas ou pas suffisamment présent dans le corpus d'entraînement. On peut donc décrire l'algorithme comme suit :

$$p(w_i|w_{i-1}) = \begin{cases} P(w_i|w_{i-1}) & \text{Si } c(w_i|w_{i-1}) > 0 \\ \alpha\, p(w_i) & \text{Autrement} \end{cases}$$

où α est le poids du repli, qui est choisi pour s'assurer que $\sum_{w \in V} p(w_i) = 1$ où w est un mot compris dans le vocabulaire V.

Pour résumer, si un trigramme n'est pas disponible en quantité suffisante (ici, au moins une fois), on se repliera alors sur le bigramme et finalement sur l'unigramme qui lui doit être présent puisque le mot fait partie du vocabulaire.

2.4.2 Lissage (Smoothing)

Le lissage est une technique couramment utilisée en modélisation statistique du langage. L'idée générale du lissage est celle-ci : retirer une masse de probabilité aux événements observés pour la redistribuer sur des événements qui ne l'ont pas été. Au fil des années, plusieurs méthodes ont été proposées pour estimer la probabilité des événements rarement ou jamais observés. Une étude [Chen 1996] exhaustive a d'ailleurs été réalisée afin de comparer chacune de ces méthodes dans la pratique.

Pour évaluer le maximum de vraisemblance d'un *n*-gramme, l'estimateur utilisé est la fréquence relative. Le maximum de vraisemblance, comporte un biais auquel les différentes techniques de lissage tentent d'apporter une réponse. Ce biais a pour effet de donner trop d'importance aux événements observés et de discréditer ceux qui ne l'ont pas été.

Dans les sections qui suivent, chacune des idées soutenant les techniques est présentée et une formule présente la modification apportée pour obtenir les probabilités du modèle de langage résultant. Ces techniques sont présentées dans un ordre pédagogique, permettant au lecteur de se familiariser d'abord avec les techniques les plus simples pour ensuite présenter des techniques moins intuitives qui sont, néanmoins, celles utilisées dans le domaine afin de produire les meilleurs modèles de langage.

2.4.2.1 Lissage Plus1 (Add-one smoothing)

Le lissage Plus1 est la technique la plus simple et elle est basée sur la loi de Laplace datant de 1814. Laplace considère que si un événement ne s'est pas produit mais qu'il est probable qu'il se produise, il devrait être vu au moins une fois. Malheureusement, cette idée fonctionne mal dans la pratique puisqu'elle consiste à prétendre que tous les *n*-grammes ont été vus une fois de plus qu'en réalité. Ceci pose un problème important : étant donné le nombre très grand de *n*-grammes qui peuvent se manifester, elle offre une trop grande masse de probabilité aux événements non vus et sous-estime les événements qui l'ont été.

La probabilité de cette technique de lissage se calcule comme suit :

$$p_{plus1}(w_i|w_{i-1}) = \frac{1 + c(w_{i-1}, w_i)}{|V| + c(w_{i-1})}$$

2.4.2.2 Lissage additif (Additive smoothing)

Cette technique implique une réorganisation de la distribution des probabilités par rapport à la méthode Plus1. Elle est basée sur les travaux de [Lidstone 1920] et ceux de [Jeffreys 1948]. Comme Plus1 accorde trop d'importance aux événements non vus, on va plutôt ajouter un nombre fractionnaire pour modifier la probabilité des *n*-grammes. Bien que cette technique soit plus efficace que Plus1, ses performances sont quand même décevantes car elle aussi, donne une importance trop grande aux *n*-grammes non vus.

$$p_{add}(w_i|w_{i-1}) = \frac{\delta + c(w_{i-1}, w_i)}{\delta|V| + c(w_{i-1})} \quad ,0 < \delta < 1$$

2.4.2.3 Estimateur de Good-Turing et Lissage de Katz

L'estimateur de Good-Turing [Good 1953] est une technique de lissage qui s'attaque aux *n*-grammes peu fréquents. Cependant, elle n'est pas utilisée directement pour le lissage de *n*-grammes car elle n'inclus pas la combinaison des modèles d'ordres supérieurs avec ceux d'ordres inférieurs, nécessaire pour obtenir une bonne performance. C'est avec la technique de lissage de Katz [Katz 1987] que la combinaison des ordres est faite.

Pour tout *n*-gramme rencontré r fois, on transforme sa fréquence en $r^* = (r+1)\frac{n_r + 1}{n_r}$ où n_r est le nombre de *n*-grammes apparus r fois. On applique cette technique sur les *n*-grammes dont la fréquence est plus faible étant donné que leurs estimateurs sont peu fiables contrairement aux événements rencontrés plus de r fois qui sont considérés, eux, suffisamment bien estimés par la technique du maximum de vraisemblance. De cette manière, l'algorithme de Katz donne une meilleure probabilité aux *n*-grammes qui ont de meilleurs (*n-1*)-grammes.

L'algorithme de Katz se présente comme suit :

$$d_r = \frac{\frac{r^*}{r} - \frac{(k+1)n_{k+1}}{n_1}}{1 - \frac{(k+1)n_{k+1}}{n_1}}$$

où $d_r = 1$ pour tous les *n*-grammes présents plus de k fois. Pour les *n*-grammes vus moins de k fois, d_r se rapproche de 1 à mesure que la valeur de k augmente. Généralement, on applique cette modification aux *n*-grammes dont la fréquence est d'environ 7 [Katz 1987]. La probabilité se calcule de la manière suivante :

$$p_{Katz}(w_i|w_{i-1}) = \begin{cases} C(w_{i-1}w_i)/C(w_{i-1}) & \text{Si } r > k \\ d_r C(w_{i-1}w_i)/C(w_{i-1}) & \text{Si } k \geq r > 0 \\ \alpha(w_{i-1})p(w_i) & \text{Si } r = 0 \end{cases}$$

avec

$$\alpha(w_{i-1}) = \frac{1 - \sum_{w_i:r>0} p_{katz}(w_i|w_{i-1})}{1 - \sum_{w_i:r>0} p(w_i)}$$

2.4.2.4 Witten-Bell

Pour la technique de Witten-Bell [Witten 1991], le coefficient servant à modifier les fréquences des *n*-grammes ne dépend pas du nombre de fois qu'un *n*-gramme a été observé mais plutôt du nombre d'observations C différentes qui se sont produites à la suite d'un contexte particulier.

$$p_{WB}(w_i|w_{i-1}) = \frac{c(w_{i-1}, w_i)}{c(w_i) + C}$$

Dans le cas des unigrammes, C correspond tout simplement à la taille du vocabulaire.

2.4.2.5 Lissage Absolu

Pour cette technique [Ney 1994], une constante b est soustraite à la fréquence de chacun des *n*-grammes.

$$p_{abs}(w_i|w_{i-1}) = \frac{c(w_{i-1}, w_i) - b}{c(w_i)}$$

où b est souvent défini comme la borne supérieure $b = \frac{n_1}{n_1 + 2n_2}$ et n_1 et n_2 correspondent respectivement au nombre d'observations qui sont survenues exactement 1 et 2 fois. Cette technique a pour effet de retirer davantage de probabilités aux observations peu fréquentes alors que celles qui sont fréquentes (à partir d'une fréquence de b) seront elles, favorisées.

2.4.2.6 Kneser-Ney

La technique Kneser-Ney a été développée par Kneser et Ney [Kneser 1995]. S'inspirant de la technique du lissage absolu présentée précédemment, elle s'inspire également du lissage de Witten-Bell car elle se base sur l'historique en plus des fréquences des mots. Cette technique qui semble être la plus performante à ce jour [Chen 1996]. Le n-gramme reçoit une probabilité qui n'est pas proportionnelle à sa fréquence mais plutôt au nombre de mots différents qui le suivent dans le corpus d'entraînement.

$$p_{KN}(w_i, w_{i-1}) = \begin{cases} \frac{\max\{c(w_{i-1}, w_i) - D, 0\}}{c(w_{i-1})} & \text{Si } c(w_i, w_{i-1}) > 0 \\ \gamma(w_i) p_{KN}(w_i) & \text{Si } c(w_i, w_{i-1}) = 0 \end{cases}$$

où $P_{KN}(w_i) = C(\bullet\, w_i) / \sum_{w_i} C(\bullet\, w_i)$ où $C(\bullet\, w_i)$ est le nombre de mots différents qui précèdent w_i. De plus, $\gamma(w_i)$ est choisi pour faire en sorte que la distribution somme à 1. Quant à D, c'est le même coefficient qu'aurait calculé l'algorithme de lissage absolu.

2.5 Méthodes d'élagage (cutoff et pruning)

Lorsque l'on créé des modèles de langage à base de n-grammes, il est souvent impossible, étant donné la grande quantité de n-grammes, d'utiliser ces modèles tel quels dans un système en reconnaissance de la parole. Il faut par conséquent élaguer les modèles, ce

qui imposent deux contraintes : maximiser la performance (minimiser la perplexité) et minimiser la taille[1] du modèle. Deux techniques sont discutées : le seuillage sur la fréquence et l'élagage par mesure d'entropie relative.

2.5.1 Seuillage sur la fréquence (cutoff)

Avec cette technique, on élimine tous les *n*-grammes qui n'ont pas une fréquence suffisamment importante. Il est courant d'utiliser un seuil différent pour chacun des ordres de *n*-grammes. En plus d'être intuitive, cette technique est directement liée avec l'approche du maximum de vraisemblance.

En plus, il a été montré dans [Seymore 1996a] qu'il est intéressant de diminuer la taille d'un modèle de langage en analysant le besoin que les *n*-grammes fassent partie du modèle. Ainsi, lorsque le *n*-gramme n'est pas présent dans le modèle, le principe veut qu'on utilise alors le repli au lieu de cette probabilité. Lorsque la probabilité du *n*-gramme et du repli correspondant est très proche, on peut alors laisser cette probabilité de côté et n'utiliser que le repli (qui représente davantage de *n*-grammes) pour obtenir pratiquement la même probabilité. Malgré l'élégance de cette technique, la diminution de la taille des modèles ainsi élagués n'est pas très importante mais elle peut être utilisée de manière complémentaire avec la technique d'élagage courante.

2.5.2 Par mesure d'entropie relative (pruning)

Cette technique a été introduite par [Stolcke 1998]. Au lieu de se baser uniquement sur les comptes de *n*-grammes, on s'intéresse plutôt à l'augmentation de la perplexité qu'un *n*-gramme peut causer. Si ce dernier fait augmenter la perplexité de plus qu'un certain seuil, on l'élimine. Le travail a permis d'ailleurs de comparer cette technique avec la technique

---

[1] La taille du modèle de langage est définie par le nombre d'arcs dans le transducteur le représentant. La taille T en arcs peut être approximée par sa borne supérieure $T = N_n + \sum_{i=1}^{n-1} 2N_i$ où N_i est le nombre de n-grammes d'ordre i et n est l'ordre maximal du modèle. Dans ce cas, l'ordre le plus élevé n'a pas de repli et le plus grand nombre d'arcs se produirait dans le cas où chacun des *n*-grammes d'ordre inférieur aurait une probabilité de repli.

courante qui est d'imposer une présence minimale en termes de fréquence (voir 2.5.1) des n-grammes. Malgré que l'élagage par le critère d'entropie relative soit plus efficace, ces deux techniques éliminent dans 85% des cas, les mêmes n-grammes.

2.6 Modèle avec cache

Les modèles avec cache [Kuhn 1990] sont des modèles qui appliquent une modification localisée sur les probabilités de certains n-grammes lorsqu'ils sont précédés de mots particuliers. Le cache permet d'incorporer une mémoire à court terme (généralement entre 100 et 1000 mots) dans un modèle de langage. [Clarkson 1997] conclut qu'avec une longueur d'historique de 500 mots, une amélioration de 14% de réduction de la perplexité a été obtenue.

L'estimateur cache se calcule comme suit :

$$p_{cache}(w_i|h) = \frac{1}{K} \sum_{j=i-K}^{i-1} \in (w_j = w_i)$$

où K représente le nombre de mots dont tient compte l'historique h (les mots précédents) du cache et $\in (w_j = w_i)$ est une fonction qui retourne 1 ou 0 selon que si $w_j = w_i$ ou non.

Le modèle exploitant un cache est particulièrement intéressant pour une tâche comme la dictée parce que dans un système où des nouveaux mots sont ajoutés avec peu d'exemples, ces nouveaux mots n'ont pas un estimateur construit à partir du même corpus. On suppose alors que la probabilité de ces mots est différente des statistiques globales. Ces mots ont d'ailleurs tendance à être répétés plusieurs fois dans un même contexte, suffisamment rapproché dans le temps.

2.7 Techniques d'adaptation

L'adaptation est un processus en modélisation statistique du langage qui permet d'obtenir un modèle de langage avec peu de données. Pour se faire, on utilise un modèle de langage

général avec des paramètres bien estimés. On créé également, à l'aide de peu de données, un plus petit modèle de langage, plus proche de la tâche [Federico 1996]. L'opération d'adaptation consiste en bout de course à utiliser les estimateurs des deux modèles de langage pour en construire un troisième qui sera le modèle adapté.

Dans le contexte de ce document, on s'intéresse particulièrement à deux techniques. D'abord l'interpolation linéaire qui se démarque par sa simplicité d'implantation et par son utilisation répandue et ensuite, MDE, une technique basée sur des mesures d'entropie.

2.7.1 Interpolation linéaire

L'interpolation linéaire est une technique d'adaptation qui permet de combiner plusieurs modèles de langage à partir de poids qui sont appliqués sur toutes les probabilités des *n*-grammes.

$$p(w|h)=\sum_{i=1}^{k}\lambda_i p_i(w|h) \text{ avec } \sum_{i=1}^{k}\lambda_i=1$$

où k est le nombre de modèles combinés. Cette méthode d'interpolation est la plus simple. Il est aussi facile d'implanter un algorithme basé sur EM (Expectation Maximization) afin de trouver les poids qui produiront les modèles les plus performants à partir d'une hypothèse de départ (les poids initiaux).

2.7.2 MDE (Minimum Discriminant Estimation)

MDE est une technique qui utilise un modèle d'adaptation contenant seulement les unigrammes. Elle fût introduite par [Kneser 1997a]. Dans ce contexte d'adaptation, le modèle à adapter est le modèle de base et le modèle d'unigrammes est construit pour représenter les données d'adaptation. Cette technique a pour avantage d'être particulièrement rapide sur les modèles de langage avec repli. En effet, si on considère l'ensemble τ, représentant l'ensemble des *n*-grammes observés dans les données du modèle de base, on

obtient ce qui suit pour la probabilité du mot w étant donné l'historique h dans le modèle adapté :

$$p_{adapt}(w|h) = \begin{cases} \dfrac{\alpha(w)}{z_0(h)} p_{base}(w|h) & \text{Si } (h,w) \in \tau \\ \dfrac{1}{z_1(h)} p_{base}(w|\hat{h}) & \text{Autrement} \end{cases}$$

où $\hat{h}$ est l'approximation de l'historique h lors du repli. Les fonctions de normalisation z_0 et z_1 sont définies comme suit :

$$z_0(h) = \frac{\sum_{w:(h,w)\in\tau} \alpha(w) p_{base}(w|h)}{\sum_{w:(h,w)\in\tau} p_{base}(w|h)} \text{ et } z_1(h) = \frac{1 - \sum_{w:(h,w)\in\tau} p_{adapt}(w|\hat{h})}{1 - \sum_{w:(h,w)\in\tau} p_{base}(w|h)}$$

où $\alpha(w)$ est le facteur utilisé pour optimiser les fonctions. [Kneser 1997a] le définit comme suit :

$$\alpha(w) = \left(\frac{p_{adapt}(w)}{p_{base}(w)} \right)^{\beta}$$

À l'aide d'expériences dans [Kneser 1997a], un minimum a été déterminé empiriquement. La valeur optimale suggérée pour β est 0.5.

Cette technique a été implantée dans un système supervisé dans [Niesler 2002]. Il est aussi possible d'opérer cet algorithme en mode non supervisé en déterminant les poids d'interpolation avec un algorithme itératif et un texte de développement. Elle se trouve à être particulièrement utile lorsque peu de données sont disponibles pour réaliser une adaptation étant donné que les seules informations qui sont utilisées sont les unigrammes.

2.8 Choix des techniques d'amélioration

Pour améliorer un modèle de langage, plusieurs techniques sont disponibles. Pour arriver à de bons résultats (à une amélioration significative), encore faut-il encore que ces méthodes soient appropriées aux données par lesquelles le modèle de langage sera évalué.

Les éléments à considérer pour la conception d'un modèle de langage :

1. Taille du corpus

   On pourrait argumenter qu'il n'y a jamais autant de données de disponibles qu'on pourrait en avoir besoin. En effet, des centaines de millions de mots sont généralement nécessaires pour construire un modèle de langage adéquat avec des trigrammes [Seymore 1996b]. Évidemment, les données disponibles se rapportant à la tâche d'intérêt sont généralement disponibles en quantité limitée et ajouter n'importe quels textes au corpus d'entraînement peut s'avérer inutile voir dévastateur. Il faut par conséquent bien choisir ces nouvelles données. Et puisque l'ajout de données implique une opération d'adaptation, il est préférable de connaître les particularités des algorithmes disponibles afin d'utiliser ces derniers judicieusement.

2. Augmenter l'historique

   Une solution intéressante pour augmenter la performance est d'augmenter la longueur de l'historique auquel les n-grammes s'appliquent. Cependant, pour utiliser un ordre plus élevé, il faut qu'une quantité suffisante de données soit disponibles. Il est également nécessaire de considérer les besoins en temps d'exécution mais surtout en mémoire qui augmente rapidement avec l'ordre des n-grammes.

3. Rapprocher le style du corpus au style de la tâche

   Dans le cas où le matériel d'entraînement n'est pas suffisamment proche de la nature des textes qui sont utilisés dans la tâche, il devient efficace de reformater des parties de corpus pour les rapprocher du style propre à la tâche. Par exemple, dans une tâche de reconnaissance de la parole spontanée, si le matériel de base consiste en des transcriptions de parole lue, on peut émuler certains événements propres au style spontané en ajoutant ou en modifiant des transcriptions.

4. Regrouper les informations par domaines [Seymore 1997]

Dans le cas où plusieurs sujets sont abordés dans la tâche visée, on peut regrouper les textes d'entraînement par sujets et ainsi construire des modèles de langage dépendants du domaine. Dans certains cas, les textes sont déjà classifiés par sujets rendant directement possible la construction de modèles. Dans d'autres cas, il est nécessaire d'utiliser un classificateur automatique pour obtenir une quantité de textes suffisante. En effet, à cause du grand nombre de mots nécessaires pour entraîner un modèle de langage, il est impensable d'effectuer la classification manuelle de tous les textes d'un corpus.

2.9 Conclusion

Ce chapitre a présenté une revue sur la théorie associée à la modélisation statistique du langage. La perplexité y a été présentée comme métrique pour l'évaluation des modèles de langage et ses inconvénients ont été présentés. Les modèles de langage n-grammes, les plus couramment employés en reconnaissance de la parole ont été décrits. Le problème de manque de données a été également adressé et des techniques ont été présentées afin de répondre à ce problème. Ces techniques ont d'ailleurs fait l'objet de discussions en ce qui a trait à leur performance ainsi qu'à leurs différences respectives. De plus, des techniques d'adaptation des modèles de langage ont été présentées. Ce chapitre conclut en dressant les grandes lignes relatives à la construction des modèles de langage ainsi qu'au prétraitement des données qu'il est possible de faire afin d'uniformiser le corpus de textes qui sert à l'entraînement des modèles.

3. LA RECHERCHE D'INFORMATIONS ET LA CLASSIFICATION

3.1 La recherche d'informations

Les modèles de langage et la recherche d'informations ont commencé une alliance à la fin des années 90. Il est courant de voir diverses sources, comme le Web [Zhu 2001], utilisées pour créer et améliorer des modèles de *n*-grammes automatiquement. La sélection des informations utilisées pour la mise à jour des modèles de langage est d'ailleurs d'une importance cruciale [Klakow 2000]. L'adaptation des modèles peut alors se faire par domaine, à l'aide d'un système pouvant mettre à jour des modèles de langage thématiques automatiquement [Seymore 1997].

Ce chapitre présente les techniques propres au domaine de la recherche d'informations. D'abord, des techniques de groupage sont présentées afin de classer des documents par sujet dans le but de créer des modèles de langage thématiques. Ensuite, une métrique est introduite afin d'évaluer la performance des classificateurs ainsi créés.

3.2 Techniques de groupage (Clustering)

Il a été montré [Chen 1998] que la réorganisation du corpus d'entraînement par domaines permet de diminuer la perplexité. Cependant, comme une grande quantité de textes est nécessaire pour construire des modèles de langage fiables, il faut par conséquent utiliser une technique de classification automatique afin de construire des corpus d'adaptation par domaine d'une grandeur suffisante.

Dans [Iyer 1996, Clarkson 1997], on suggère d'utiliser une technique d'adaptation itérative, où les données ont été partitionnées en de petites grappes par sujets utilisées pour l'adaptation. Dans ce cas, on utilise beaucoup (plusieurs dizaines) de domaines. Évidemment, l'avantage d'utiliser un grand nombre de domaines est qu'on peut distinguer d'une manière plus précise les domaines reliés à un texte. Par contre, dans un contexte de reconnaissance de la parole, il est nécessaire que les modèles soient chargés en mémoire et

par conséquent, un nombre limité de modèles peut être utilisé. En effet, ces techniques suggèrent de combiner les modèles entre eux, d'opérer une mixture.

Au fil des années, plusieurs techniques ont été proposées afin de regrouper l'information par sujets. Dans ces techniques, peu ou pas de travail de la part d'un être humain est nécessaire. C'est ce qui différentie d'ailleurs une tâche non supervisée d'une tâche supervisée. Dans le cadre de cette étude, on s'intéresse particulièrement aux techniques qui permettent de reproduire une classification faite par un humain et c'est pourquoi l'approche supervisée a été retenue. En effet, avec ce genre de classification, on s'assure alors qu'un utilisateur sera en mesure de choisir facilement un domaine en fonction du sujet d'intérêt.

Les techniques présentées dans ce chapitre ont l'avantage d'identifier automatiquement ces mots clés mais ont aussi également la flexibilité nécessaire pour octroyer un poids à chacun des mots contenus dans un texte. La contribution de chacun des mots d'un document produira une distance par rapport à chacun des domaines. On peut ensuite décider si c'est seulement le meilleur document qui est associé au domaine ou si ce sera les N meilleurs domaines. C'est cette association de documents en corpus par domaines qui permet de créer des modèles de langage thématiques.

3.2.1 TF-IDF (Term Frequency – Inverse Document Frequency)

Cette technique a été originalement développée dans les laboratoires d'AT&T [Salton 1991]. Elle constitue une métrique reconnue dans le domaine de la recherche d'informations. Elle consiste à établir une métrique pour chacun des mots d'un ensemble de documents qui permet de déterminer son importance à l'intérieur d'un ensemble regroupant des textes du même sujet.

Commençons par définir $tf(d_i, w_j)$, le nombre de fois qu'un mot w_j est présent dans le document d_i. Ensuite, $idf(w_j)$ qui se calcule comme suit :

$$idf(w_j) = \frac{D}{nombre\ de\ documents\ contenant\ w_j}$$

où D est le nombre de documents dans l'ensemble d'entraînement. De cette manière, $idf(w_j)$ est grand pour les mots qui sont dans peu de documents. Puis, $tfidf(d_i, w_j)$ d'un document d_i et d'un mot w_j se définit comme suit :

$$tfidf(d_i, w_j) = tf(d_i, w_j) \log(idf(w_j))$$

Ainsi, $tfidf(d_i, w_j)$ sera grand pour les mots fréquemment rencontrés dans d_i mais non dans les autres documents. Le travail effectué par cette technique est donc d'identifier les mots qui représentent les caractéristiques particulières à un sujet.

Pour effectuer la classification, on définit la similarité $S(d_i, d_k)$ entre un document d_i et d_k comme suit :

$$S(d_i, d_k) = \frac{\sum_{j=1}^{V} tfidf(d_i, w_j) \cdot tfidf(d_k, w_j)}{\sqrt{\left(\sum_{j=1}^{V} tfidf(d_i, w_j)^2\right) \cdot \left(\sum_{j=1}^{V} tfidf(d_k, w_j)^2\right)}}$$

où V représente la taille du vocabulaire. On peut également construire un vecteur $t(d_i)$ qui donne pour un document d_i, pour chacun des mots du corpus, une valeur TFIDF. Avec ce vecteur, on dispose alors de la distance moyenne qu'existe entre chacun des documents des ensembles et du document à classifier. Une approche moins gourmande consiste à calculer un vecteur moyen pour chacun des sujets pour ensuite calculer la distance qu'existe entre le document à classifier et ce vecteur moyen.

Comme mentionné dans [Niesler 2002], on peut exprimer la similarité comme le cosinus entre les deux vecteurs. De cette façon, $S(d_i, d_k)$ s'écrit aussi :

$$S(d_i, d_k) = \frac{t(d_i) \bullet t(d_k)}{\|t(d_i)\| \cdot \|t(d_k)\|}$$

où l'opération effectuée au numérateur est un produit scalaire.

Cette technique est un standard dans le domaine. En plus d'être utilisée comme base afin de comparer de nouveaux algorithmes, elle est utilisée dans plusieurs systèmes réels [Craven 2000].

3.2.2 SVM (Support Vector Machine)

La théorie derrière les SVM a été développée par Vapnik [Vapnik 1995] pour répondre aux problèmes de la reconnaissance de formes (Pattern Recognition), de régression ainsi que pour la classification. Particulièrement dans le cas de la classification de textes, l'implantation de la théorie proposée par Vapnik a été intégrée dans une suite d'outils nommée SVM^{light} [Joachims 2002]. Ce logiciel permet de créer un classificateur et ainsi de classifier n'importe quelle sorte de données représentées sous forme de vecteurs.

Bien que cette technique repose sur une théorie rigoureuse et que l'implantation dont il est question plus haut soit plus que solide, cette technique s'avère difficile à utiliser dans la pratique. Il est en effet nécessaire de fournir des exemples négatifs, difficiles à trouver, pour chacun des ensembles, le choix de ces exemples influençant grandement la performance du classificateur [Yu 2002].

3.2.3 Classification à l'aide de la perplexité

Cette technique consiste à utiliser des modèles de langage statistiques afin de construire un classificateur. Dans un premier temps, des textes sont choisis manuellement pour construire les modèles de départ, un modèle par domaine. Ces modèles sont ensuite utilisés pour classer tous les textes du corpus : la métrique de distance entre un texte et un domaine est la perplexité du texte sur le modèle du domaine.

Plusieurs paramètres ont été étudiés au fil des années avec ce type de classificateur. D'une part, l'ordre du modèle de langage qui permet une classification optimale et aussi, la technique d'association des textes aux domaines. Des études [Kneser 1997b, Gildea 1999, Mahajan 1999] ont conclu que le type de langage le plus approprié pour cette tâche était l'unigramme.

La tâche d'apprentissage est basée sur EM (Expectation Maximization) [Neal 1998]. Au fil des itérations, la perplexité décroît jusqu'à ce qu'elle arrive à un minimum. Lorsque ce minimum est atteint, on considère alors que les modèles du classificateur ont convergés. Cette convergence est importante car elle démontre que le classificateur est conséquent d'une itération à l'autre. Bien que l'algorithme EM ne trouve qu'un minimum local, il est généralement suffisant pour obtenir une classification adéquate. Le domaine qui obtient la perplexité la plus faible pour un texte donné se retrouve à être enrichi de ce texte. Itérativement, ce processus termine lorsque la perplexité globale cesse de diminuer.

3.3 Évaluation de la classification

Deux valeurs indiquent la performance d'un classificateur : le rappel et la précision [Jones 1981]. Le rappel indique la proportion des résultats pertinents dans la base de données qui ont été retournés par la requête. La précision quant à elle indique la proportion des résultats retournés qui sont pertinents à la requête.

Posons A comme étant l'ensemble des éléments pertinents et B l'ensemble des résultats retournés. Également, les ensembles a,b,c et d sont décrits dans la figure 2. Le rappel et la précision sont définis comme suit :

$$\text{rappel} = P(B|A) = \frac{P(A \cup B)}{P(A)}$$

$$\text{précision} = P(A|B) = \frac{P(A \cup B)}{P(B)}$$

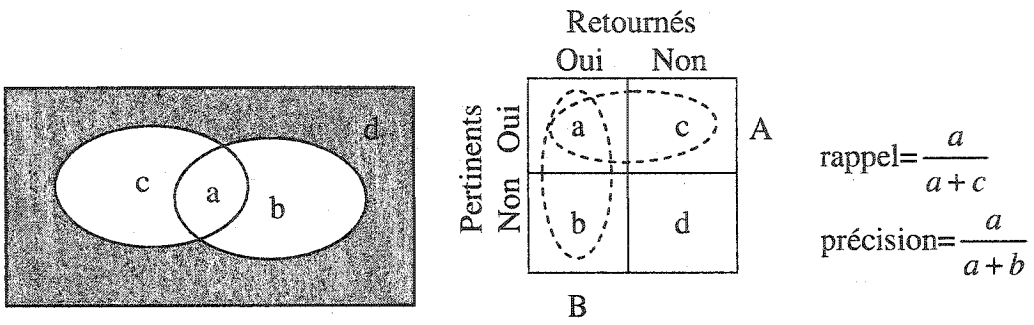

Figure 2: Calcul de la performance d'un classificateur

Le nombre de documents pertinents correspond donc à $a+c$ et le nombre de documents retournés par le requête est $a+b$. Ces deux mesures complémentaires sont donc nécessaires pour l'évaluation d'une requête.

Bref, un haut taux de rappel indique que peu de documents pertinents ont été laissés de côté. En revanche, un taux de précision élevé quant à lui indique que les résultats retournés sont pertinents dans une proportion élevée. Cette mesure à deux volets est utile pour bien cerner la performance d'un classificateur. Par exemple, dans le cas d'un classificateur avec un haut taux de fausses acceptations, ce classificateur serait alors caractérisé par un taux de rappel élevé et une précision basse.

Enfin, cette technique d'évaluation est dépendante de la requête qui est effectuée. Pour évaluer un ensemble de classificateurs où le meilleur gagne par exemple, le rappel et la précision globale seront les mêmes car tous les textes sont nécessairement associés à tort ou à raison, à une classe. Cette valeur correspond à la performance globale du classificateur.

3.4 Conclusion

Ce chapitre a introduit les domaines de recherche que sont la recherche d'information ainsi que la classification de documents. Après avoir effectué un bref survol de la création de modèles de langages thématiques grâce à la classification de documents, trois méthodes ont été revues. Particulièrement, la technique de classification à l'aide de la perplexité sera celle qui sera utilisée dans les prochaines sections de ce document. La technique TF-

IDF quant à elle servira de référence afin de valider nos choix et afin de s'assurer d'une performance minimale. Enfin, les notions de rappel et de précision ont été introduites afin de quantifier la qualité des classificateurs.

4. CONTEXTE DE RECHERCHE

Le système de reconnaissance automatique de la parole du CRIM [Boulianne 2000] peut avoir plusieurs utilités. Il a déjà été utilisé pour post-synchroniser des films [Boulianne 2003] et pour effectuer de la dictée vocale. Cette fois-ci, c'est dans un contexte impliquant de la parole spontanée à grand vocabulaire et une contrainte de produire une transcription raisonnable en temps réel dans un système de sous-titrage chez un télédiffuseur que le défi se présente [Brousseau 2003].

Dans le projet couvert par ce document, c'est particulièrement l'actualité traitée par les médias d'information qui sont au coeur de l'étude. L'utilisation des supports informatiques est maintenant chose courante depuis les dix dernières années chez la totalité des grands télédiffuseurs. Il est normal qu'un système voulant produire des sous-titres à l'aide de la reconnaissance de la parole vienne se greffer à un tel système pour en extraire toutes les informations pertinentes. Évidemment, l'utilisation d'Internet tend à faciliter l'accès aux informations tout autour de la planète. Un tel système permet donc d'adapter les modèles de langage à partir des nouvelles sur les sites Internet, des fils de presse et surtout, de la feuille de route disponible en partie sur les serveurs informatiques utilisés pour réaliser les bulletins de nouvelles.

Ce chapitre présente le système auquel cette étude s'intéresse. On présente d'abord comment les modèles de langage sont crées. Ensuite, on présente les techniques qui ont été retenues pour la mise à jour de ces modèles.

4.1 Présentation du système

À partir d'une référence de base, plusieurs idées ont été envisagées pour augmenter le taux de reconnaissance dont l'une consiste à utiliser plusieurs modèles de langage, orientés vers les sujets de discussion dans les domaines de l'actualité traités. C'est ce problème qui est décrit dans ce chapitre.

Voici comment le système fonctionne :

1. L'utilisateur (le perroquet[2]) sélectionne quel domaine il souhaite utiliser avant le début de la séance de travail ;

2. Il répète mais aussi rephrase en simultané ce qu'il entend puis la sortie du reconnaisseur est utilisée pour produire les sous-titres.

La figure 3 illustre l'architecture utilisée dans le cadre de ce projet.

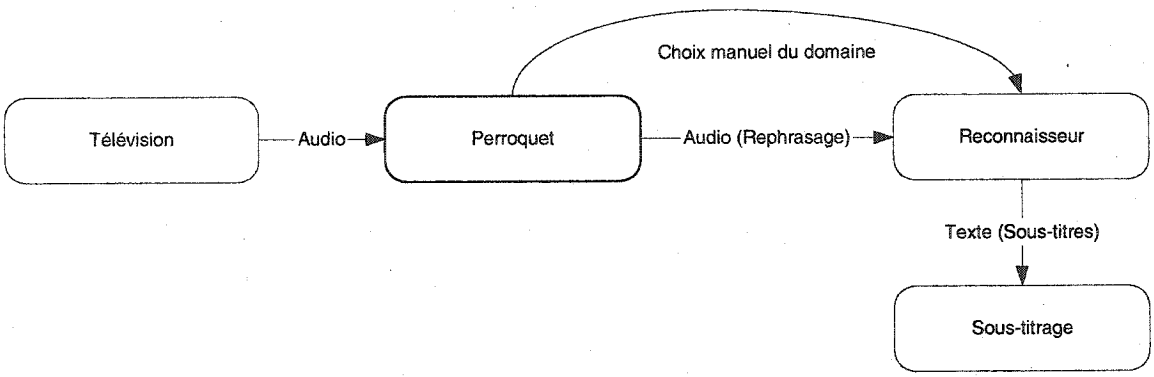

Figure 3: Architecture du système de sous-titrage

4.2 Corpus de textes

Afin de créer des modèles de langage, deux sources de données étaient disponibles soit les trois dernières années disponibles sur les serveurs informatiques de TVA et une base de données de 1999 à 2003 des journaux québécois : La Presse et Le Soleil, provenant de la compagnie CEDROM-SNi.

Le nombre de mots pour l'entraînement de chacun des domaines et de chacun des sources est présenté dans la table 7.

[2] Le terme perroquet (ou respeaker) décrit la tâche de l'utilisateur qui est de répéter ce qu'il entend.

Domaines	Nombre de mots TVA	Nombre de mots CEDROM-Sni	Nombre de mots total
Le Monde	182.7k	9.3M	9.5M
National	526.6k	17.9M	18.4M
Régional	422.7k	16M	16.4M
Culture	596.4k	18.6M	19.2M
Économie	114.4k	12.9M	13M
Sports	206.5k	11M	11.2M
Météo	20.5k	600k	620.5k

Table 7: Tailles des corpus utilisés pour réaliser les expériences

4.3 Création des domaines

Dans l'optique de créer des modèles de langage thématiques, c'est-à-dire dépendants du domaine, il est nécessaire de choisir des sources de textes dont le contenu soit suffisamment proche des nouvelles qui seront d'intérêt lors de la réalisation du bulletin de nouvelles. En effet, on s'attend à utiliser des sources qui sont susceptibles de contenir des informations reliées aux sujets traités lors de la mise en ondes.

Dans le cadre particulier de ces travaux, on s'intéresse aux textes de nouvelles disponibles en ligne. Que ce soit des fils de presse ou des textes de journaux disponibles sur Internet, l'identification des sources à utiliser s'avère d'une importance particulière car l'adaptation des modèles de langage qui en découlera pourrait bien, si les sources ne sont pas adéquates, dégrader la performance de la reconnaissance.

De plus, le style particulier des textes de nouvelles est à prendre en considération. La couverture des nouvelles, autant géographique que les sujets traités a également une

grande importance. En effet, un site de nouvelles de la France sera moins pertinent pour un système s'intéressant à des nouvelles du Canada et du Québec comme un site spécialisé sur la finance sera peu utile si les nouvelles financières ne sont pas traitées dans le bulletin de nouvelles.

Afin d'améliorer le taux de reconnaissance, on s'intéresse à segmenter un modèle de langage en plusieurs modèles qui seront plus pointus sur certains domaines de l'actualité. Mais comment choisir ces domaines ? Combien en choisir ? Toutes ces questions ont trouvé une réponse à l'aide des critères suivants :

- la possibilité qu'un être humain puisse reproduire cette classification,
- qu'il soit possible à un être humain de choisir rapidement quel domaine est le meilleur,
- minimiser la superposition,
- ressembler le plus possible à ce qui existe déjà comme classification.

Les sites de nouvelles utilisés sont présentés à la table 8.

Source de données	Sujets traités
Canoë http://www.canoe.qc.ca	Circulation, Faits divers, Le Monde, Météo, National, Régional, Montréal, Culture, Économie, Techno
RDI http://ici.radio-canada.ca/nouvelles	Culture, Sports, Météo, Le Monde, Politique, Économie, Santé et environnement, Internet et médias, Montréal
La Presse http://cyberpresse.ca	Actualités, Arts et spectacles, Hobbies et loisirs, Internet, Le Monde, Politique, Sciences, Sports

Table 8: Les sources de données et les sujets traités

À partir de ces informations et des sujets traités à l'intérieur d'un bulletin de nouvelles chez TVA, 7 domaines ont été retenus. Ils sont présentés à la table 9.

Nom du domaine	Sujets traités
Le Monde	L'actualité internationale, la guerre
National	L'actualité du Canada et du Québec, le parlement, la politique nationale
Régional	L'actualité des régions, les affaires municipales, les faits divers
Culture	Les arts et spectacles, les acteurs de cinéma, les critiques littéraires
Économie	Les finances, les investissements
Sports	Les sports professionnels (hockey) et amateurs (olympiques)
Météo	Les humeurs de dame nature, les conversations sur la météo

Table 9: Les domaines retenus et les sujets traités

Ces domaines répondant aux critères énoncés précédemment, c'est eux qui ont été retenus pour la création de modèles de langage thématiques. Il est facile pour un être humain d'identifier rapidement à quel domaine une nouvelle est la plus proche.

4.4 Mise à jour des modèles

Une fois ces modèles construits, leurs mises à jour permettra de conserver une performance acceptable. La mise à jour des modèles est effectuée à l'aide de nouveaux documents de nouvelles, traitant de l'actualité qui sera débattue en ondes. L'actualité étant en constante évolution, les modèles de langage d'un système de reconnaissance de la parole se doivent, eux aussi, d'évoluer dans le temps pour être constamment à jour avec les nou-

velles de dernières heures. Il est par conséquent nécessaire que les modèles puissent obtenir les informations pertinentes et récentes, leur permettant de refléter ces changements.

4.4.1 Adaptation à court terme

C'est avec les textes des nouvelles récentes que cette adaptation est effectuée. On peut utiliser des poids afin de pondérer [Kobayashi 1998] chaque jour. On suggère dans un tel modèle d'utiliser une pondération plus importante pour les nouvelles les plus récentes. L'utilisation de transcriptions de bulletins de nouvelles précédents a donc également été mise à contribution afin de diminuer la perplexité et d'augmenter le taux de reconnaissance d'une manière plus appréciable.

Ces transcriptions peuvent être obtenues de deux manières. La première technique, non supervisée, consiste à utiliser les sorties précédentes du système de reconnaissance. La seconde, supervisée, plus coûteuse en temps, consiste à corriger ces sorties à la main.

4.4.2 Adaptation à long terme

Une fois que suffisamment de nouveaux textes ont été accumulés, on peut penser remplacer les modèles de base par des modèles mis à jour. Il a d'ailleurs été montré que la technique MAP [Janiszek 2000] permet de réaliser ce type d'adaptation. Les nouveaux modèles pourront alors être, à leur tour, adaptés, jour après jour. Il est alors important de bien combiner ces textes avec le modèle de base. Malgré l'importance de cette tâche, cette étude se concentrera particulièrement sur la mise à jour à court terme.

4.4.3 Traitement des nouveaux mots

Dans un contexte en constante évolution, le vocabulaire, comme les modèles, seront appelés à changer donc à s'adapter. Comme il existe un impact réel et non négligeable entre le nombre de mots hors vocabulaire et le taux d'erreur [Pallet 1994], il est donc important d'ajouter les mots qui sont susceptibles d'être utilisés pendant la reconnaissance. D'un autre côté, on ne veut pas non plus ajouter n'importe quel mot car on risquerait alors de

dégrader la performance des modèles [Rosenfeld 1996]. Ainsi, un choix judicieux doit être fait quant à l'inclusion des nouveaux mots à ajouter au vocabulaire.

La plus simple des techniques pour ajouter un nouveau mot est de ne l'ajouter que sous la forme d'un unigramme et de leur octroyer la probabilité du mot inconnu. Dans la littérature [Jelinek 1990] on propose un modèle se basant sur des mots synonymes, définis comme étant les mots qui sont rencontrés dans un même contexte. Des techniques plus évoluées existent comme par exemple la méthode d'adaptation MDE [Kneser 1997a] qui donne la probabilité des unigrammes du modèle d'adaptation au modèle adapté. C'est d'ailleurs cette technique qui sera utilisée pour l'expérimentation pour la mise à jour des modèles de langage ainsi que pour l'ajout des nouveaux mots.

Dans la recherche de nouveaux mots, il est possible que certains documents contiennent des mots erronés dus à des erreurs de normalisation. Comme certaines sources de données peuvent se retrouver sur Internet et qu'elles sont par conséquent incontrôlables, il faut également prévoir un mécanisme pour prévenir que des mots indésirables viennent contaminer les modèles. Comme un tel système ne peut être parfait, on peut utiliser la connaissance des utilisateurs pour valider la liste de nouveaux mots à vérifier. Les mots inappropriés peuvent être de cette façon écartés.

4.5 Utilisation d'un modèle générique

Dans le cas où un modèle thématique n'est pas suffisamment performant ou qu'il ne se rapproche pas suffisamment du domaine d'une nouvelle, on peut utiliser un modèle de langage générique. L'idée d'utiliser un modèle générique vient un peu briser le modèle de l'application qui utilise des modèles thématiques mais en ayant recours à un modèle qui peut être appliqué à toutes les situations, on s'assure en fait d'une qualité minimale. En effet, on peut utiliser le modèle générique de deux façons : la première, on l'interpole avec les modèles thématiques, la deuxième, on utilise le modèle générique indépendamment (un modèle de plus) lorsque le sujet traité dans la tâche n'est pas relié spécifiquement aux domaines thématiques disponibles.

Il est également intéressant d'utiliser un modèle générique dans le cas où l'utilisateur n'est pas en mesure d'identifier à l'avance quel sera le sujet abordé dans une section. Il est en effet probable que la performance du modèle générique soit plus élevée en effectuant une tâche quelconque qu'en utilisant un modèle thématique sur un domaine avec lequel ce dernier n'est pas suffisamment lié de près.

4.6 Conclusion

Ce chapitre a couvert le contexte de la recherche qui a entouré la recherche effectuée dans cette étude. D'abord, le système de sous-titrage a été présenté. Ensuite, les corpus de textes disponibles pour la construction initiale des modèles thématiques sont décrits. Puis, les détails de la création ainsi que les techniques de mises à jour ont été présentés afin d'adresser deux problématiques pour la mise à jour des modèles de langage soit la mise à jour à court terme (journalière) et le traitement des nouveaux mots.

5. EXPÉRIENCES ET RÉSULTATS

Ce chapitre présente le cheminement ainsi que différents résultats qui ont été obtenus suite aux travaux réalisés dans le cadre d'un projet exécuté au Centre de recherche informatique de Montréal. Ce projet avait pour but de produire des sous-titres à l'aide de la technologie de reconnaissance automatique de la parole du CRIM pour les émissions en direct du réseau TVA.

Dans un premier temps, une série d'expériences a permis d'évaluer l'impact de l'utilisation des modèles de langage thématiques par rapport à une approche ne comportant qu'un seul domaine : le modèle générique. Dans un deuxième temps, d'autres expériences sont présentées sur la manière de mettre à jour les modèles de langage thématiques automatiquement.

5.1 Étude sur les modèles de langage thématiques

La suite d'expériences de cette section s'intéresse à l'utilisation de modèles de langage thématiques pour améliorer le taux de reconnaissance. Premièrement, la construction manuelle des classificateurs y est présentée. Ensuite, l'évaluation des modèles thématiques y est faite. Les expériences de cette section ont été réalisées sur les parties en direct du bulletin de TVA de 22h du 24 septembre 2002.

5.1.1 Création des modèles thématiques

Tout d'abord, ce sont les textes de La Presse (Montréal) et du Soleil (Québec) qui ont été extraits à partir d'une banque de textes achetée chez CEDROM-SNi. Ces textes, étant en format XML, comportent déjà plusieurs champs de classification qui indiquent à quels sujets un article est relié. De cette façon, les quelques 430 sujets relevés dans cette banque de données ont été manuellement associés aux 7 domaines retenus pour notre classificateur. Les modèles de langage ont été créés à l'aide de SRILM [Stolcke 2002]. Les modèles de langage ont été créés avec la technique de lissage Witten-Bell et avec un vocabulaire pour chaque domaine d'environ 20000 mots.

La table 10 donne le nombre de sujets avant[3] et après l'extraction des textes, le nombre de sujets dans la colonne « Avant » précise le nombre de sujets qui ont été manuellement retenus pour la construction des modèles au départ et le nombre de sujets dans la colonne « Après » indique le nombre de sujets qui sont finalement présents une fois que l'extraction des textes de la base de données est complétée.

Domaines	Avant	Après
Le Monde	24	389
National	73	426
Régional	66	425
Culture	25	384
Économie	75	408
Sports	3	307
Météo	11	200

Table 10: Nombre de sujets après et avant l'extraction

La table 11 fait l'état de la superposition entre les domaines par rapport aux sujets avant et après l'extraction des textes. Pour lire la première ligne par exemple, « 15 » « 388 », cela signifie que les domaines « National » et « Régional » partage 15 sujets choisis manuellement et qu'une fois tous les textes extraits, en reprenant la liste des sujets de tous les ensembles, ils en partagent 388. Il y avait en tout 433 sujets dans la banque de textes de CEDROM-SNi.

[3] La classification qui à été utilisée pour chacun des domaines est disponible à l'annexe A.

Avant/Après	Monde		National		Régional		Culture		Économie		Sports		Météo	
Monde														
National	**15**	388												
Régional	6	387	**55**	424										
Culture	0	357	0	381	1	379								
Économie	1	374	6	403	3	402	3	374						
Sports	0	296	0	306	0	305	0	296	1	299				
Météo	0	197	0	200	0	199	1	357	0	374	0	178		

Table 11: Superposition avant et après l'extraction

La superposition « avant » montre une superposition plus forte pour les domaines Monde-National et pour National-Régional ; qui se justifie facilement étant donné que ces domaines traitent de nouvelles semblables mais à des échelles différentes. Pour ce qui est de la deuxième colonne (superposition « après »), elle montre clairement qu'une fois tous les textes des domaines rassemblés et la liste des sujets extraits de nouveau, le champ « sujet » n'est plus fiable pour une reclassification des textes étant donné le côté pointu de ce champ.

Une fois ces textes mis ensembles pour créer un modèle de langage propre au domaine, il a été nécessaire d'adapter ces domaines au style de TVA. Pour ce faire, tous les textes disponibles (c'est-à-dire 3 ans) provenant du serveur de mise en ondes des bulletins ont été utilisés pour se rapprocher du style journalistique de TVA. Les articles ont été classifiés à l'aide du classificateur se basant sur la perplexité développé dans le cadre de ce travail.

Voici un tableau qui résume la perplexité des modèles de langage par rapport aux ensembles de développement :

Dom / Dév	Le Monde	National	Régional	Culture	Économie	Sports	Météo
Le Monde	29	128	138	178	158	163	114
National	91	80	80	113	100	97	89
Régional	128	102	48	133	118	122	113
Culture	112	108	95	85	100	104	90
Économie	118	113	97	135	68	118	103
Sports	123	121	109	130	127	57	100
Météo	115	123	109	137	121	117	57

Table 12: Perplexité des modèles par domaines sur tous les ensembles de développement

À nouveau, la perplexité est résumée ici, par rapport au modèle de langage générique, puis le modèle sur son propre ensemble de développement :

Dev.	Le Monde		National		Régional		Culture		Économie		Sports		Météo	
Ppl (générique et thématique)	96	29	99	80	87	48	115	85	97	68	92	57	88	57

Table 13: Comparaison de la perplexité du modèle générique et par domaines

La perplexité d'un texte de développement est donc la plus basse lorsque mesurée sur le modèle de langage s'y rapportant. Ceci montre donc que les modèles thématiques ont appris adéquatement des caractéristiques de chacun des domaines.

5.1.2 Évaluation du classificateur

Pour réaliser un test sur la performance du classificateur, 50 articles ont été retenus pour faire partie d'un ensemble de test. La table 14 présente l'erreur de classification produite par les domaines. « 1er sujet choisi » indique que le sujet retenu était le meilleur, « 2 meil-

leurs » que le sujet retenu était dans les 2 meilleurs et « 3 meilleurs » que le sujet retenu était parmi les 3 meilleurs domaines.

	Erreur
1er sujet choisi	22/50 (44%)
2 meilleurs	9/50 (18%)
3 meilleurs	9/50 (18%)

Table 14: Erreurs de classification (résultats bruts)

Pour donner un sens réel à ce test, il faut cependant retirer quelques articles qui contiennent peu ou pas d'information qui aurait permis à un être humain d'arriver à la même conclusion. La table 15 montre ce résultat, avec 3 articles de moins qu'au départ :

	Erreur
1er sujet choisi	19/47 (40%)
2 meilleurs	6/47 (13%)
3 meilleurs	4/47 (9%)

Table 15: Erreurs de classification (résultats fins)

La table 16 présente le nombre d'erreurs et le nombre d'exemples disponibles pour chacun des domaines :

Domaine	Nombre d'erreurs / Nombre de vidéos
Monde	5/8
National	7/12
Régional	4/10
Culture	0/11
Économie	0/2
Sports	0/1
Météo	3/3

Table 16: Détails des erreurs pour chacun des domaines

La figure 4 présente en détails la performance individuelle de chacun des classificateurs des domaines (la performance globale du classificateur est de 60%):

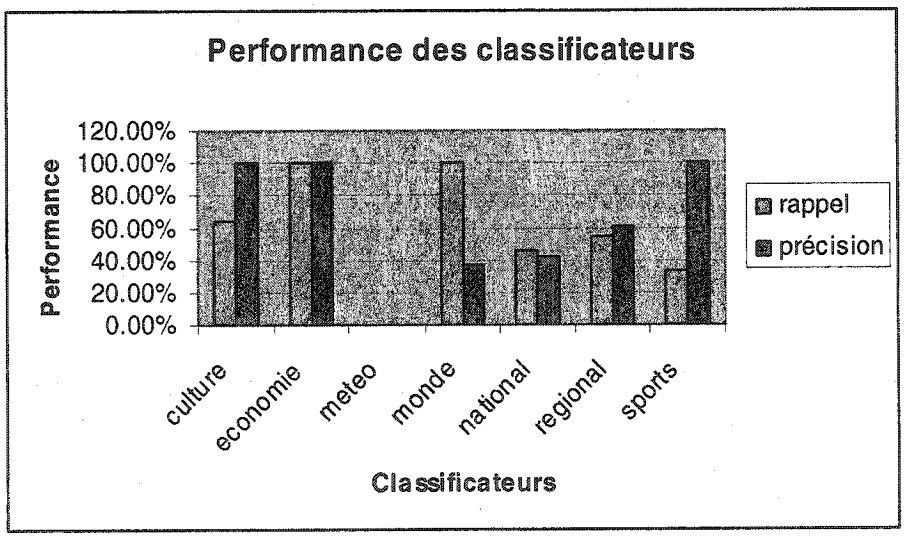

Figure 4: Performance des classificateurs

À la lumière de ces résultats, on peut conclure que la performance du domaine « Météo » n'est pas adéquate. En effet, le classificateur de ce domaine ne réussit pas à classer aucun de ses propres exemples. Au contraire, le domaine « Culture » fonctionne très bien mais il y a un grand nombre de fausses acceptations d'où la proportion de rappel basse. De plus, comme il existe une superposition entre les domaines « Monde », « National » et « Régional », ces trois domaines ont tous une proportion de rappel basse. Il en est égale-

ment de même pour le domaine « Sports ». La construction manuelle de ce classificateur ne semble pas être tout à fait au point.

5.1.3 Évaluation des modèles de langage par domaine

L'évaluation des modèles de langage par domaine est en effet une question qu'il est nécessaire de soulever pour s'assurer qu'en bout de ligne, le seul véritable critère qui nous intéresse : le taux de reconnaissance, se montre satisfaisante.

La table 17 présente le taux de reconnaissance du modèle générique par rapport aux modèles de langage les plus performants sur un premier ensemble d'enregistrements. Dans ce tableau, ce sont les résultats du meilleur domaine qui sont reportés pour chacun des vidéos, à la manière d'un oracle qui connaît à l'avance quel sera le meilleur domaine pour une nouvelle. Ces performances représentent la limite supérieure de la performance de l'approche par domaines, présentée ici.

Meilleurs modèles (oracle) / Modèle générique			
No	Titre du vidéo	Domaines	Générique
1	au retour, discours trône	84.78%	76.09%
2	enfant mort – intro	76.64%	63.55%
3	À venir, est-ce la fin MG	80.56%	69.44%
4	irak, vite des inspecteurs	56.25%	54.69%
5	après pause, drogue	75.76%	63.64%
6	fonctionnaires victimes	84.44%	73.33%
7	médecins - connai. dossiers	83.23%	71.26%
8	opinions par email	84%	65%
9	Dumont – intro	77.06%	74.31%
10	discours trône - topo	87.5%	75%
11	Dumont - discussion	79.61%	76.7%
12	dernière réunion des minis.	86.67%	80.83%
13	discours trone - intro	78.87%	76.06%
14	désintoxication - intro	93.75%	82.81%
15	médecins - venus en nombre	78.01%	72.25%
16	médecins	63.85%	63.85%
17	finance – bourse	92.82%	84.1%
18	finance - massif de P.R.S.F	91.67%	85.42%
19	météo	57.77%	57.77%
20	marc gagnon - va à retraite	86.67%	80%
21	chrétien vs irak	72.41%	55.17%
22	israël + cisjordanie	47.37%	26.32%
23	irak - nucléaire dans 2 ans	64.71%	56.47%
24	médecins et loi 114 - intro	85.9%	79.49%
25	mère qui bat ses enfants	67.47%	66.27%
26	irak avec Blair – intro	75%	72.44%
27	miss univers	63.64%	48.25%
	Moyenne	**76.90%**	**68.54%**

Table 17: Comparaison du taux de reconnaissance des modèles par domaines et du modèle générique

Ces résultats sont reproduits dans ce graphique. On peut y remarquer que le taux de reconnaissance de l'approche par domaine est systématiquement plus élevé que celui du modèle générique. Les mêmes résultats sont présentés graphiquement dans la figure 5 suivant le même ordre que la table 17.

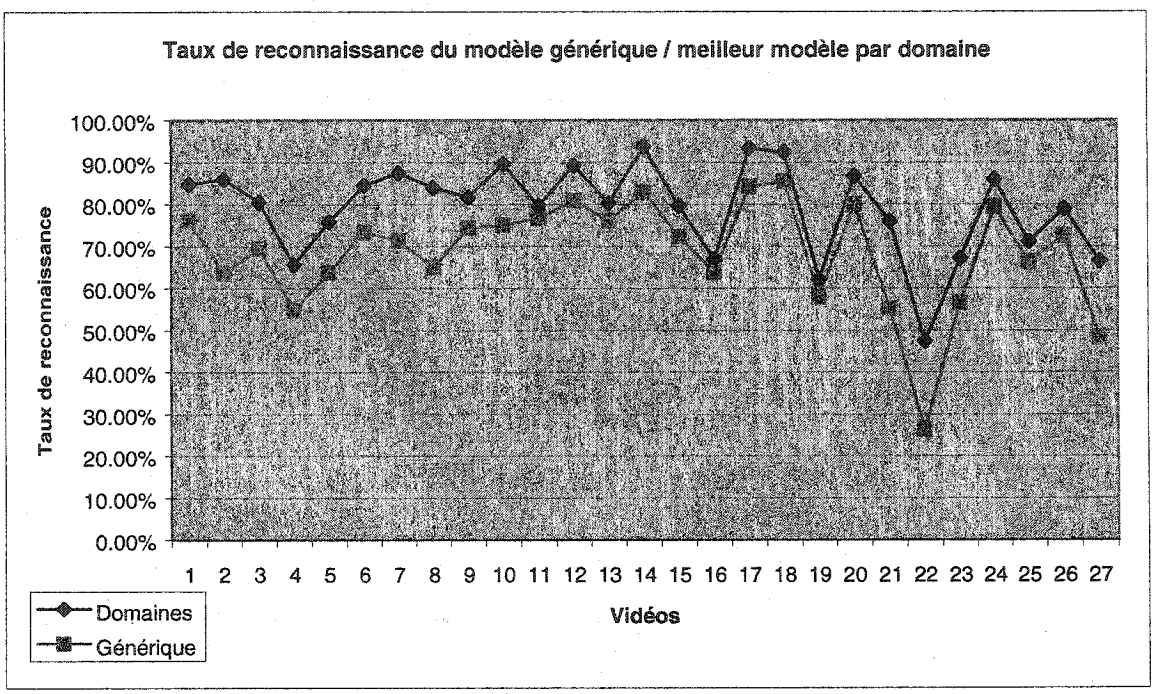

Figure 5: Taux de reconnaissance du modèle générique / meilleur modèle par domaine

Bien que ces résultats soient satisfaisants (une augmentation en moyenne du taux de reconnaissance de 8.36% absolu), ils ne représentent pas ce qui se produirait dans la réalité. En fait, dans une application en utilisation réelle, ce n'est pas toujours le meilleur domaine qui serait choisi.

Comme le classificateur est supposé reproduire le comportement humain pour la classification par domaine de nouvelles, c'est exactement ce qui sera fait ici. La table 18 introduit cette idée. La transcription des vidéos est utilisée pour classer celui-ci et ainsi, choisir le modèle de langage le plus approprié. La colonne « Vidéos » identifie sommairement chacun des vidéos (présentés dans le même ordre dans le graphique précédent), la colonne « Domaines » identifie avec quel domaine a été retenu manuellement, la colonne « Retenu » est la performance du domaine choisi, la colonne « Rang » indique à quel rang le domaine retenu se trouve dans une liste de domaines triée par taux de reconnaissance pour ce vidéo, la colonne « Perte » précise la perte engendrée par cette erreur de classification (colonne « Retenu » moins le taux de reconnaissance à la colonne « Domaines »

dans le tableau précédent) et la colonne « Générique » indique, à titre de référence, la performance du modèle générique sur ce vidéo.

Résultats en choisissant le domaine manuellement						
No	**Vidéos (en ordre)**	**Domaines**	**Retenu**	**Rang**	**Perte**	**Générique**
1	au retour, discours trône	national	80.43%	1	0%	76.09%
2	enfant mort – intro	régional	70.09%	2	6.55%	63.55%
3	à venir, est-ce la fin MG	sports	80.56%	1	0%	69.44%
4	irak, vite des inspecteurs	monde	53.12%	3	3.13%	54.69%
5	après pause, drogue	régional	75.76%	1	0%	63.64%
6	fonctionnaires victimes	régional	84.44%	1	0%	73.33%
7	médecins - connai. Dossiers	national	83.23%	1	0%	71.26%
8	opinions par email	culture	84%	1	0%	65%
9	dumont – intro	national	71.56%	4	5%	74.31%
10	discours trône – topo	national	87.50%	1	0%	75%
11	dumont – discussion	national	77.67%	2	1.94%	76.70%
12	dernière réunion des minis.	national	79.17%	2	8%	80.83%
13	discours trone – intro	national	70.42%	6	8.45%	76.06%
14	désintoxication – intro	national	73.44%	6	20%	82.81%
15	médecins - venus en nombre	national	78.01%	1	0%	72.25%
16	médecins	régional	63.85%	1	0%	63.85%
17	finance – bourse	économie	92.82%	1	0%	84.10%
18	finance – massif de P.R.S.F	économie	86.11%	5	6%	85.42%
19	météo	météo	57.77%	1	0%	57.77%
20	marc gagnon - va à retraite	sports	86.67%	1	0%	80%
21	chrétien vs irak	monde	62.07%	2	10.34%	55.17%
22	israël + cisjordanie	monde	47.37%	1	0%	26.32%
23	irak - nucléaire dans 2 ans	monde	45.88%	6	19%	56.47%
24	médecins et loi 114 – intro	régional	85.90%	1	0%	79.49%
25	mère qui bat ses enfants	régional	63.86%	5	4%	66.27%
26	irak avec Blair – intro	monde	73%	2	2%	72.44%
27	miss univers	culture	63.64%	1	0%	48.25%
	Moyenne		**73.27%**	**2.222**	**3.5%**	**68.54%**

Table 18: Résultats en choisissant le domaine manuellement

5.1.4 Discussion

Les domaines permettent donc d'obtenir un gain de 4.73% (absolu) en moyenne lorsqu'un utilisateur sélectionne le domaine avant d'effectuer la reconnaissance. L'utilisateur, pour ce test, a réussi à sélectionner le domaine offrant le plus haut taux de reconnaissance presque sept fois sur dix (dans 68% des cas). Ceci montre bien que le classificateur basé

sur la perplexité réussit à reproduire le jugement humain, ce qui était un des objectifs au moment de sa conception.

Aussi, on peut remarquer que le taux de reconnaissance chute rapidement (même jusqu'à être plus bas que celui du modèle générique) lorsque le meilleur domaine n'est pas choisi. En effet, on peut observer jusqu'à 19% de dégradation dans le taux de reconnaissance. Ce comportement est dû au fait que les modèles de langage sont maintenant spécialisés, qu'ils sont plus performants dans certaines conditions particulières, mais qu'il faut les utiliser dans les bonnes situations. D'ailleurs, dans le vidéo nommé « finance – bourse », le meilleur domaine est choisi ce qui permet de faire chuter le taux d'erreur de 15% à 7%.

5.2 Étude de la mise à jour des modèles de langage

Un autre aspect intéressant à étudier est la mise à jour des modèles de langage avec les textes de nouvelles disponibles sur les fils de presse, les sites de nouvelles sur le Web ou sur tout autre support informatique disponible.

Le système étudié par les expériences décrites dans cette section est un système de recherche d'informations, de classification de documents et de mise à jour de modèles de langage. Dans un premier temps, les documents de certains sites Web ou systèmes d'informations (comme le serveur de nouvelles du télédiffuseur) sont téléchargés. Puis, à l'aide de classificateurs construits spécialement pour chaque source de données, ces documents sont classés par domaines. Finalement, les modèles de langage thématiques du système sont mis à jour à l'aide des textes reliés à chacun d'entre eux.

Les expériences ont été réalisées sur le bulletin de TVA de 17h30 du 13 novembre 2003. 47 extraits étaient en direct et ce sont eux qui ont été retenus pour le test. Les données d'adaptation ont été les sites de nouvelles de TVA, RDI et de La Presse.

Cette section aborde une autre série d'expériences sur la classification de textes de nouvelles par domaine. D'abord, de nouveaux modèles de langages thématiques ont été créés avec un classificateur qui effectue plusieurs itérations afin d'atteindre une stabilité (convergence). Ensuite, les classificateurs ainsi créés seront évalués. On abordera aussi le

traitement accordé aux nouveaux mots pour élargir les vocabulaires de chacun des domaines. De plus, l'adaptation journalière des modèles de langage sera évaluée.

5.2.1 Création des modèles thématiques

Les modèles de langage thématiques on été construits avec les mêmes sources de textes que dans la section précédente. Cependant, un algorithme EM a été utilisé pour la construction des nouveaux classificateurs afin d'améliorer la performance de classification. Ce sont les mêmes domaines qui ont été retenus que dans la section 5.1. Ici par contre, deux sortes de modèles de langage sont créés : les modèles de langage pour la classification et les modèles de langage pour la reconnaissance. Un autre ensemble de test a été créé pour l'évaluation de la mise à jour car nous ne disposions d'aucune donnée pour la mise à jour des modèles pour les expériences présentés en 5.1.

Cette section présente les étapes ainsi que les choix qui ont été effectués pour mener à la création des modèles de langage de classification. Les tests abordés dans cette section sont :

- l'étude de la variation de l'ordre des modèles de langage utilisés par le classificateur ;
- l'étude de la variation des paramètres des modèles de langage de classification, c'est-à-dire le vocabulaire ;
- la vérification de la performance des nouveaux classificateurs en la comparant à une technique déjà connue ;
- la démonstration de la convergence des modèles, preuve de la cohérence entre chaque itération.

5.2.1.1 Ordre des modèles

L'ordre des modèles de langage est l'un des premiers paramètres qu'on s'intéresse à faire varier pour étudier la performance. C'est le paramètre le plus important d'un modèle de

langage parce qu'il fera varier d'une manière importante la quantité de données qui sera nécessaire pour entraîner le classificateur. La figure 6 présente une comparaison de la performance des différents classificateurs utilisant des longueurs d'historiques différentes.

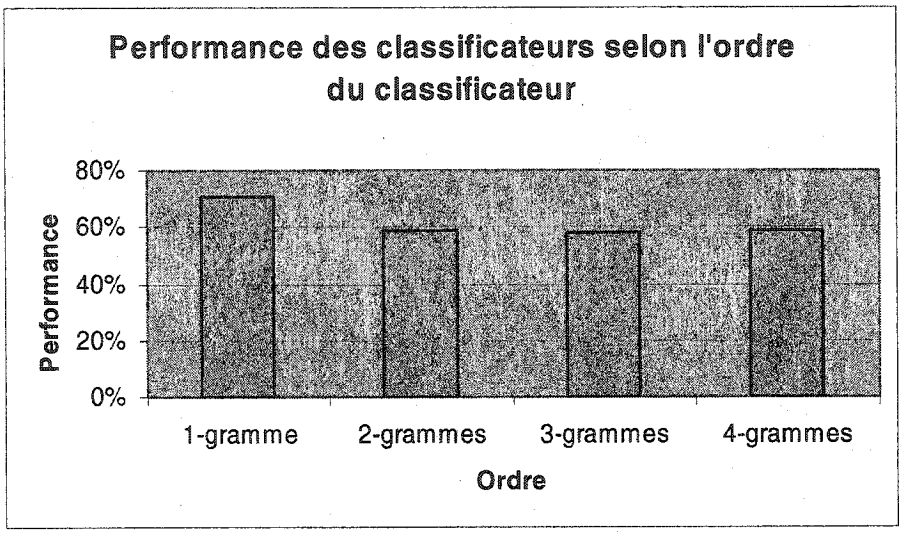

Figure 6: Performance de classification selon l'ordre du classificateur

Cette expérience montre que ce sont les modèles à base d'unigrammes qui sont les plus performants. Ce sont donc des modèles unigrammes qui seront comparés dans les prochaines sections. Ce sont en effet les modèles qui nécessitent le moins de données.

5.2.1.2 Choix des paramètres pour le classificateur

Les paramètres du classificateur étudié ici sont les mots qui sont utilisés par les modèles de langage. Lors de la construction initiale du classificateur, ces idées ont été testées :

1. Le vocabulaire des textes associés aux domaines dans l'ensemble d'entraînement

   À chaque itération, le classificateur associe un document à un domaine. Lorsque tous les documents sont associés, les paramètres de la prochaine itération correspondent au vocabulaire de chacun des domaines et le processus continue.

2. Utiliser tous les mots de l'ensemble d'entraînement

   Tous les mots de tous les documents sont considérés par les classificateurs.

3. Utilisation de TF-IDF

   Le vocabulaire des classificateurs correspond aux meilleurs mots de l'algorithme TF-IDF. Ces mots sont ceux qui ont une valeur TF-IDF de plus de 0.1 (ce qui donnait les meilleurs modèles avec cette méthode). Il y a eu en tout 3258 mots qui correspondaient à ce critère.

4. Les 20000 mots les plus fréquents dans l'ensemble d'entraînement

Les idées ont été essayées dans cet ordre. La première et la deuxième techniques donnant des résultats désastreux (moins de 20% de performance), ce sont les deux autres qui se sont avérées fonctionner pour effectuer une classification convenable.

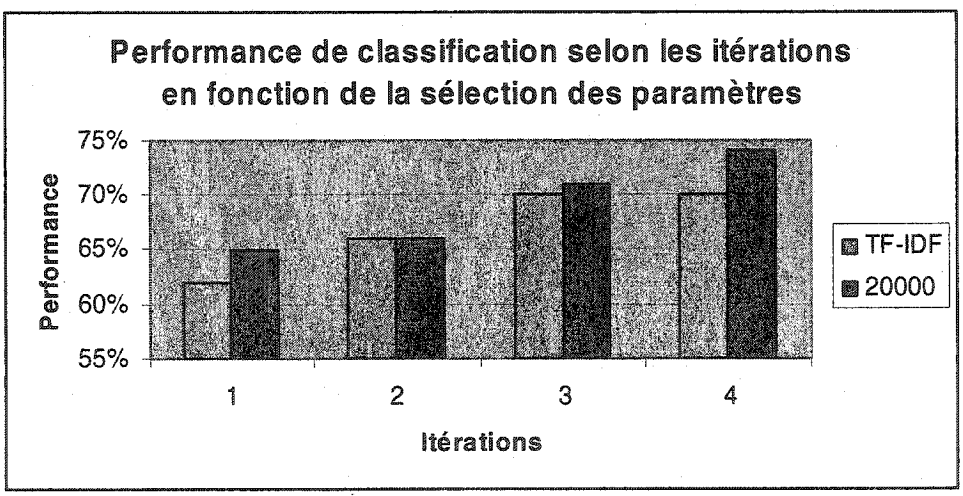

Figure 7: Performance de classification selon les itérations en fonction de la sélection des paramètres

En plus d'avoir un plus grand pouvoir de classification dès le départ, les 20000 mots les plus fréquents du corpus d'entraînement se trouvent à être faciles à générer. Les classificateurs ont donc été entraînés avec ces 20000 mots.

La comparaison de la technique TF-IDF est d'ailleurs faite à la section 5.2.1.3.

5.2.1.3 Comparaison avec TF-IDF

Afin de valider l'implantation de cette technique de classification, elle a été comparée avec une technique bien connue, TF-IDF (présentée à la section 3.2.1). D'abord, la figure 8 présente la performance du classificateur de TVA avec la méthode de la perplexité.

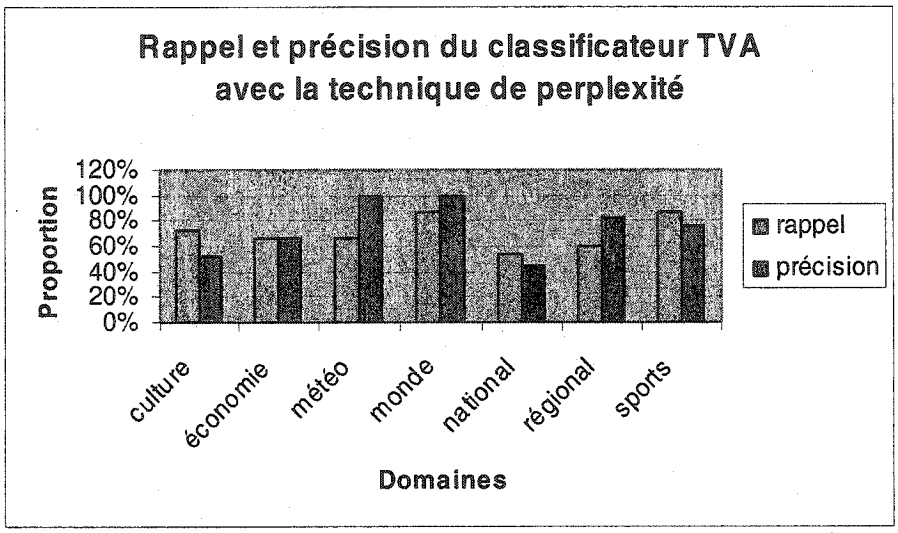

Figure 8: Rappel et précision du classificateur TVA avec la technique de perplexité

La figure 9 présente la performance du classificateur TVA implantée avec la technique TF-IDF :

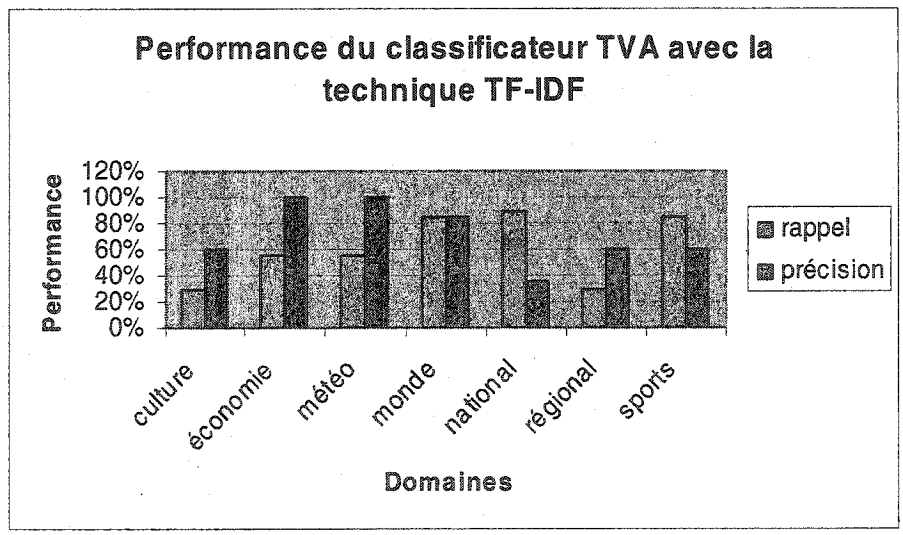

Figure 9: Performance du classificateur TVA avec la technique TF-IDF

Le classificateur à base de modèles de langage est donc plus performant que la technique TF-IDF. En effet, alors que le classificateur obtient 70% de taux de classification, la tech-

nique TF-IDF n'obtient que 61% et ce, avec les mêmes données d'entraînement. Ce résultat vient donc appuyer la technique de classification à l'aide de modèles de langage.

5.2.1.4 Convergence des classificateurs

Au cours de ces expériences, trois sources de données ont été retenues pour obtenir des nouveaux textes afin de faire la mise à jour des modèles de langage. Ces sources sont TVA, RDI et La Presse. Le choix a été fait en fonction de la performance des modèles générés avec ces sources.

Dans les figures de cette section, les barres indiquent la performance du classificateur et les courbes indiquent la perplexité des textes de l'ensemble d'entraînement. L'algorithme EM utilisé devrait en principe, à mesure qu'on procède avec les itérations, faire diminuer la perplexité sur l'ensemble d'entraînement et augmenter le taux de classification. Le processus itératif cesse lorsque la perplexité sur l'ensemble d'entraînement ne baisse plus.

Le classificateur de TVA comporte sept domaines à classifier. Après 4 itérations, il obtient une performance de 70%. La figure 10 présente les itérations menant à la convergence du classificateur TVA.

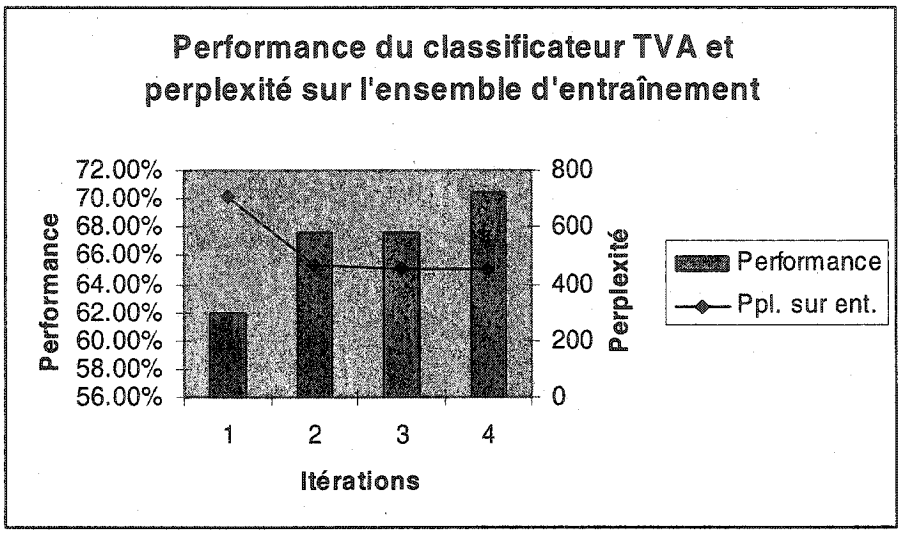

Figure 10: Performance du classificateur TVA et perplexité sur l'ensemble d'entraînement

La figure 11 illustre en détails le comportement du classificateur TVA.

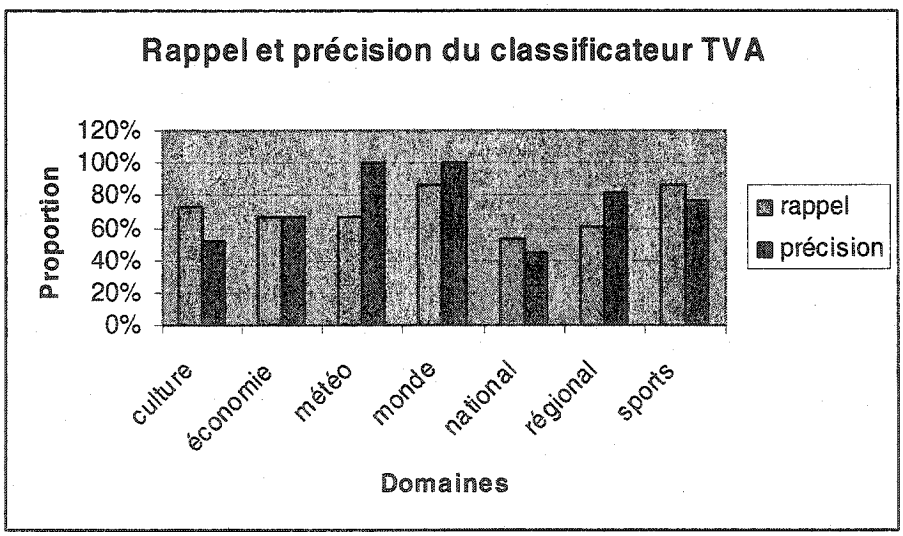

Figure 11: Rappel et précision du classificateur TVA

Le classificateur de RDI n'a pour sa part que quatre domaines à classifier. La performance globale de ce classificateur est de 91%. La performance du classificateur RDI est présentée à la section à la figure 12.

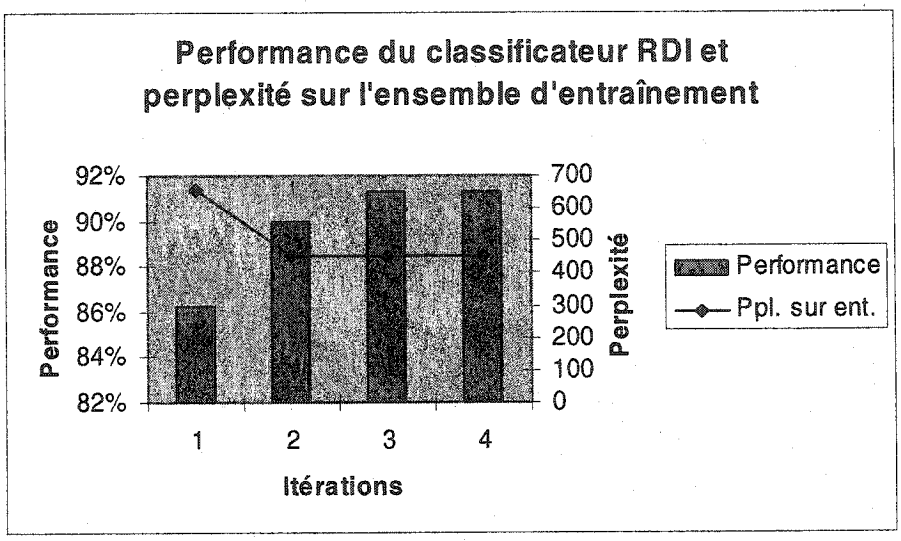

Figure 12: Performance du classificateur RDI et perplexité sur l'ensemble d'entraînement

La performance de chacun des domaines du classificateur RDI est présentée à la figure 13.

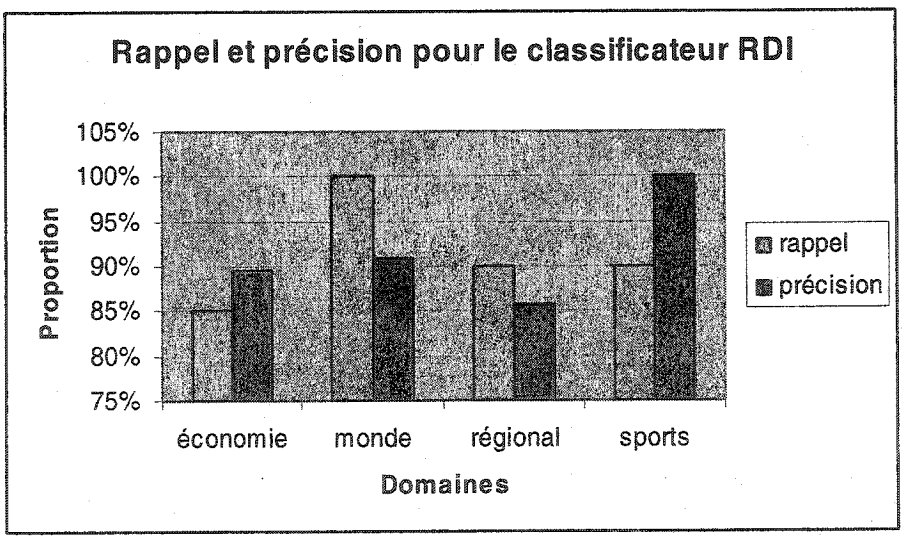

Figure 13: Rappel et précision pour le classificateur RDI

Le classificateur de La Presse a quant à lui la tâche de classifier les textes dans sept domaines. La performance globale de ce classificateur est de 70%. La performance de ce classificateur est présentée à la figure 14.

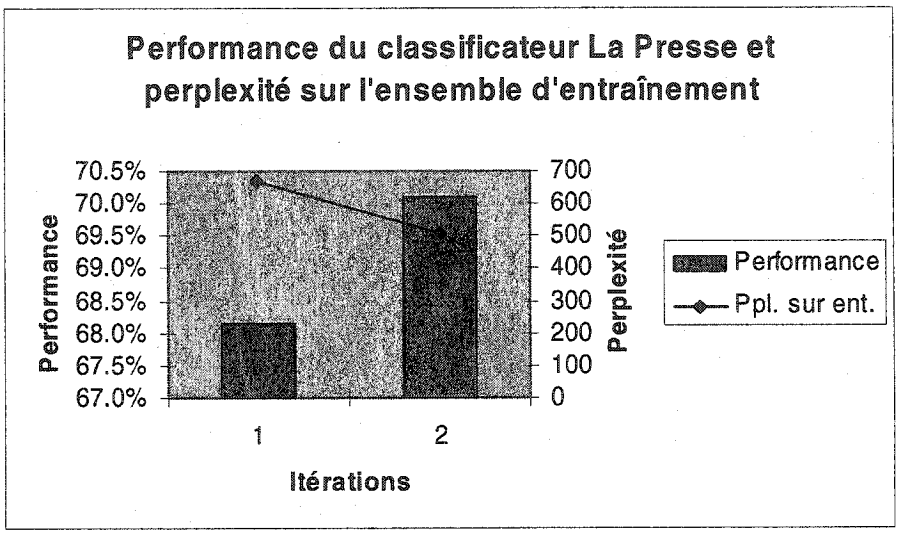

Figure 14: Performance du classificateur La Presse et perplexité sur l'ensemble d'entraînement

La performance de chacun des domaines du classificateur de La Presse est présentée à la figure 15.

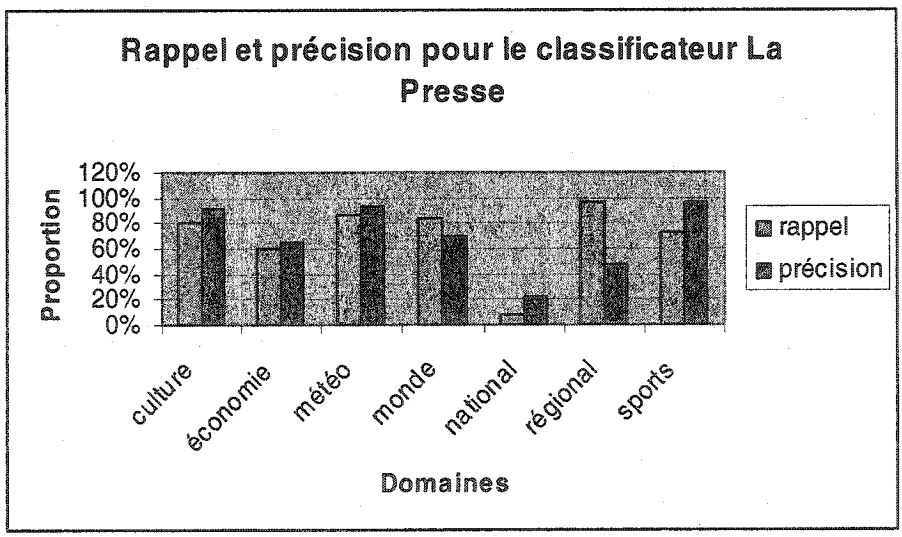

Figure 15: Rappel et précision pour le classificateur La Presse

En résumé, la construction des classificateurs est efficace en ce sens que la perplexité diminue bel et bien d'itération en itération. De plus, une augmentation du taux de classification se produit à chaque itération. La performance très élevée du classificateur de RDI

s'explique de deux façons. Premièrement, il n'y a que 4 domaines à apprendre. Deuxièmement, il n'y a pas autant de superposition car il n'y a même pas de domaine national. En effet, en plus de devoir choisir entre 7 domaines, TVA et La Presse doivent choisir entre « Régional » et « National », ce qui n'est pas chose facile, même pour un humain. C'est le domaine « National » autant pour TVA que pour La Presse qui fait surtout défaut. En effet, pour le classificateur TVA, la performance du domaine « National » se situe juste en dessous des 60% mais pour La Presse, les résultats sont vraiment décevants avec moins de 20%.

5.2.2 Sélection des nouveaux mots

Dans la section 2.3.1, il est question de la sélection du vocabulaire et de l'impact qu'ont les mots hors vocabulaire sur la reconnaissance de la parole. Un avantage d'effectuer une mise à jour des modèles de langage est qu'il est possible d'ajouter des nouveaux mots dans le vocabulaire.

Les nouveaux mots proviennent des documents associés à chacun des domaines. Les mots qui ne figurent pas alors dans le vocabulaire du domaine sont ajoutés. La figure 16 montre l'amélioration de cette mise à jour sur le pourcentage de mots hors vocabulaire.

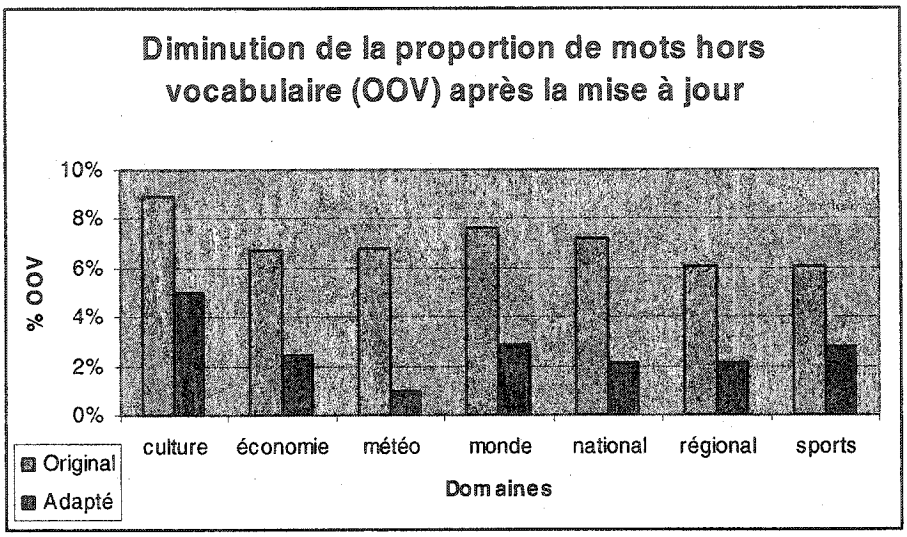

Figure 16: Diminution de la proportion de mots hors vocabulaire (OOV) après la mise à jour

Il n'y a que la mise à jour du domaine culture qui ne permette pas de diminuer le taux de mots hors vocabulaire en bas de 4%. Ceci s'explique par le fait que les nouvelles de ce domaine contenaient beaucoup de noms propres et que ces noms propres ne faisaient pas partis des textes téléchargés. C'est d'ailleurs ce domaine qui a le pourcentage de mots hors vocabulaire le plus élevé. Malgré ce problème, la diminution du pourcentage de mots hors vocabulaire est appréciable. On remarque en effet une amélioration d'environ 50% sur le pourcentage de mots hors vocabulaire.

5.2.3 Combinaison des données d'adaptation

Pour évaluer la mise à jour des modèles de langage, la technique d'adaptation a d'abord été évaluée. Le point le plus important est la combinaison des informations en fonction du temps qui les séparent de la séance de travail. On s'attend, par exemple, que les données les plus récentes se voient assigner une plus grande importance. Les résultats présentés dans la figure 17 montrent que ce sont les données de la journée en cours (jour0) qui sont les plus importants.

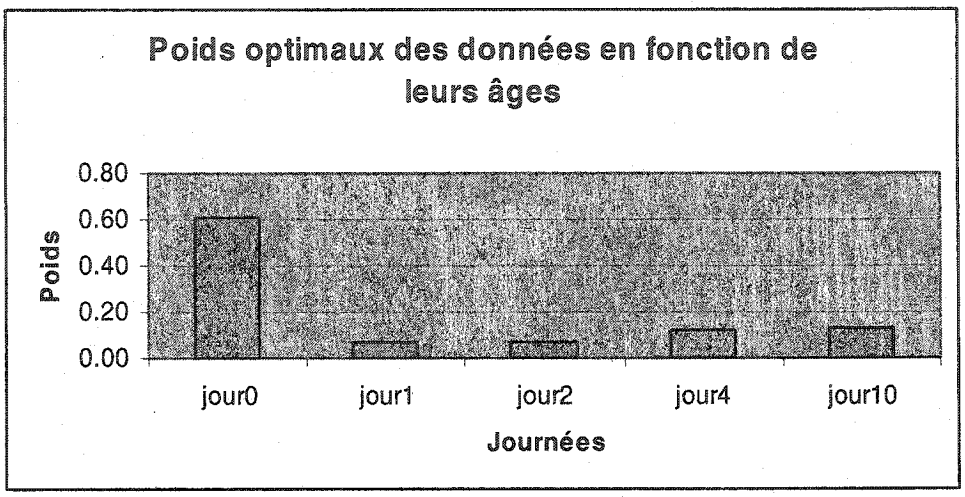

Figure 17: Poids optimaux des données en fonction de leurs âges

Pour évaluer l'importance de chacun des journées, les modèles de chacune des journées ont été combinés et les poids donnant la perplexité minimale correspondent à l'importance de la journée. Ces poids sont les poids moyens qui ont été calculés sur deux journées distinctes.

Pour l'évaluation de cette combinaison, des tests ont été effectués afin de déterminer la tendance de la perplexité sur des modèles mis à jour avec de plus en plus de données. Les résultats sont présentés à la figure 18.

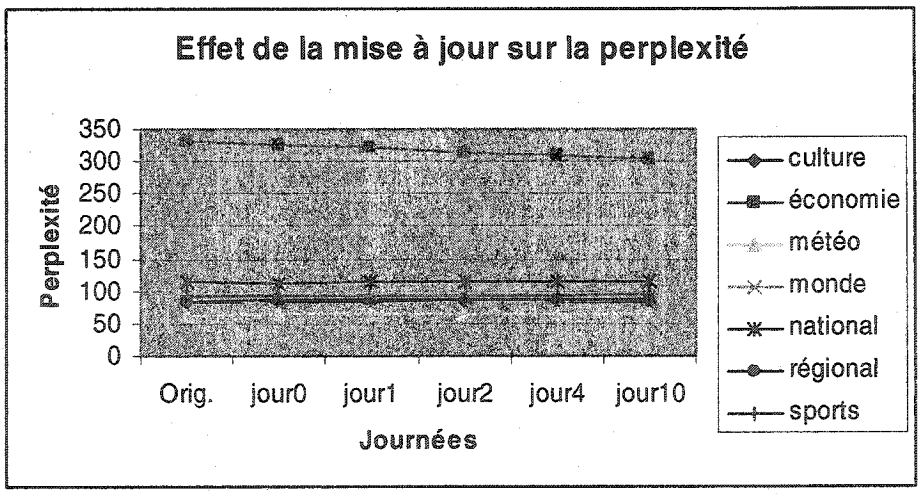

Figure 18: Effet de la mise à jour sur la perplexité

Le but de la figure 18 n'est pas de juger de l'amélioration de l'efficacité pour chacun des domaines mais bien de dégager la tendance de l'effet de la mise à jour avec les données de chacune des journées. Plus précisément, il n'y a que le domaine « Économie » qui

s'améliore d'une manière appréciable. Cependant, c'est aussi ce domaine qui a une perplexité trois fois plus élevée que les autres domaines. Pour ce qui est des autres domaines, la tendance de la perplexité sur leurs ensembles de développement respectifs est que la journée de la diffusion (jour0) est la seule qui permet l'amélioration de la perplexité. En effet, à mesure qu'on ajoute des données de plus en plus âgées, la perplexité se dégrade.

Ces résultats montrent que la perplexité se dégrade à mesure qu'on remonte dans le temps. Autrement dit, il n'y a que la journée courant (jour0) qui permet d'améliorer la perplexité ou de ne pas trop la dégrader. C'est donc les données provenant de cette journée qui ont été retenu pour les expériences suivantes.

5.2.4 Évaluation de la mise à jour

La perplexité des documents des domaines a été calculée sur chacun des modèles de langage. Il n'y a que pour le domaine « Météo » que la perplexité se dégrade d'une manière significative. Les autres domaines ont peu ou pas d'amélioration sur la perplexité. Cependant, les domaines « Économie », « Monde », « National » et « Sports » sont améliorés.

Perplexité sur la transcription des bulletins		
Domaines	Original	jour0
culture	83.1	83.3
économie	332	325
météo	62.7	65.2
monde	89	79.7
national	113.6	112.6
régional	84.8	85.5
sports	94.4	92.3

Table 19: Perplexité sur la transcription du bulletin

Étant donné que la perplexité ne permet pas de conclure que la mise à jour des modèles de langage thématique dégrade ou améliore les modèles, des expériences de reconnaissance ont été effectués.

La table 19 présente les résultats de l'expérience de reconnaissance.

Effet de la mise à jour sur le taux de reconnaissance							
No	Domaine	Original	Mis à jour	No	Domaine	Original	Mis à jour
1	régional	80.3%	81.82%	**24**	régional	64%	64%
2	sports	81.82%	81.82%	**25**	régional	65.45%	65.09%
3	régional	59.46%	59.46%	**26**	culture	79.44%	79.44%
4	météo	80%	80%	**27**	culture	66.67%	65.43%
5	régional	78.01%	78.01%	**28**	culture	65.96%	66.15%
6	sports	72.64%	72.64%	**29**	culture	70%	71.67%
7	sports	67.31%	67.31%	**30**	culture	83.33%	83.33%
8	régional	78.6%	78.23%	**31**	culture	81.13%	81.13%
9	régional	53.06%	57.14%	**32**	régional	69.49%	69.49%
10	national	83.12%	83.12%	**33**	économie	100%	100%
11	national	76%	76%	**34**	météo	82.61%	82.61%
12	national	70.37%	71.3%	**35**	régional	71.56%	72.27%
13	régional	83.41%	83.41%	**36**	sports	77.78%	78.49%
14	météo	67.35%	71.43%	**37**	sports	66.95%	67.52%
15	météo	63.89%	63.89%	**38**	national	20%	20%
16	météo	77.37%	77.37%	**39**	régional	71.01%	70.41%
17	national	96.3%	96.3%	**40**	sports	64.29%	65.71%
18	sports	67.98%	68.42%	**41**	national	56.57%	58.29%
19	régional	72.18%	71.71%	**42**	national	58.82%	58.82%
20	sports	72.03%	72.03%	**43**	culture	48.84%	46.51%
21	météo	67.6%	67.17%	**44**	régional	65.6%	66.67%
22	régional	69.23%	67.95%	**45**	météo	78.97%	78.37%
23	monde	95%	95%	**46**	national	80.95%	79.37%
				47	sports	62.5%	62.5%
		Original	Mise à jour				
	Moyenne	72.06%	72.22%				

Table 20: Effet de la mise à jour sur le taux de reconnaissance

Cette mise à jour a été effectuée à l'aide de 5000 mots provenant de textes de nouvelles des sources TVA, La Presse, RDI. Ces mots sont ceux des documents de nouvelles de la journée courante (jour0).

5.2.5 Discussion

Le taux de reconnaissance ne s'améliore en moyenne que de 0.16%. Ces résultats sont moins bons qu'on pourrait s'y attendre. En effet, avec la diminution marquée du taux de mots hors vocabulaire présentée à la section 5.2.3, l'opération de mise à jour n'apporte qu'une maigre amélioration sur le taux de reconnaissance. En revanche, ces performances s'expliquent par la qualité des données d'adaptation. Malgré ces données, une légère amélioration a été observée, ce qui confirme la pertinence d'utiliser cette technique. En effet, au lieu de rechercher des documents qui traitent des mêmes sujets que ceux qui seront

traités que dans les bulletins de nouvelles, nous avons plutôt pris tous les documents de nouvelles qui sont disponibles sur les autres sites. Par exemple, la source RDI de cette journée ne traite même pas de la nouvelles de sports d'Éric Gagné qui a gagné le trophée Cy Young. Finalement, cette expérience n'a pas permis d'évaluer la performance du domaine « Économie ». En effet, étant donné qu'il n'y a qu'un seul vidéo associé à ce domaine, on ne peut pas conclure si la performance de 100% est représentative malgré une très grande perplexité pour ce domaine. Aussi, aucune source ne contenait davantage d'informations que ce qu'avait placé le télédiffuseur sur son serveur à propos de l'entrevue sur le spectacle de « Pied de Poule » pour le domaine « Culture ».

5.3 Conclusion

Ce chapitre n'a abordé que quelques facettes du domaine de recherche qu'est la reconnaissance de la parole : la création et l'adaptation de modèles de langage thématiques. La particularité de ces expériences est que les textes qui sont utilisés pour créer et pour adapter ces modèles sont des textes qui sont classés par sujets automatiquement par un classificateur. Bien qu'il a été montré que l'utilisation des modèles thématiques permet d'obtenir une amélioration intéressante (5%) du taux de reconnaissance, la mise à jour des modèles quant à elle, ne change pas d'une manière significative la performance du système. En revanche, étant donné la qualité et la quantité d'informations disponibles pour l'adaptation lors de ces expériences, il est au moins intéressant de constater que le taux de reconnaissance se maintient. Il sera en effet nécessaire de revoir la sélection des documents pour l'adaptation des modèles de langage.

6. CONCLUSION

Ce document a couvert le projet de recherche au CRIM ayant pour but de produire des sous-titres pour les émissions de nouvelles en direct. Deux techniques ont été utilisées afin d'avoir à notre disposition de meilleurs modèles de langage pour la reconnaissance :

- la construction de modèles de langage thématiques ;
- la mise à jour de ces modèles de langage.

L'utilisation de modèles de langage thématiques a permis d'obtenir une amélioration de performance de 5% du taux de reconnaissance. La mise à jour des modèles de langage a cependant donné des résultats moins intéressants même si une amélioration minime a été obtenue après l'adaptation.

Dans ce document, il n'a pas été seulement question du côté théorique sur la modélisation du langage mais aussi des lignes directrices de la construction des modèles de langage ainsi que des choix qu'il est possible de faire afin de créer et d'adapter ces mêmes modèles. La modélisation du langage est un domaine de recherche où il reste encore beaucoup de travail à faire avant d'être en mesure de rivaliser avec la capacité d'un être humain.

6.1 Travaux futurs

À la lumière des résultats, certains points faibles ont été identifiés particulièrement sur la technique de classification des textes et sur l'opération de mise à jour des modèles. Afin de pallier aux problèmes de classification automatiques, les idées suivantes seront explorées :

- utiliser un classificateur hiérarchique afin d'obtenir de meilleures performances pour les domaines avec lesquels nous avons observés de la superposition (comme entre les modèles « Régional » et « National ») ;
- explorer des techniques supplémentaires qui n'ont pas fait l'objet de tests dans ce document comme les SVM (voir la section 3.2.2) ;
- combiner avec des techniques existantes comme TF-IDF.

Pour ce qui a trait à la mise à jour des modèles de langage, les points suivants seront mis à profits :

- utiliser un modèle de langage générique en combinaison avec les modèles thématiques afin de laisser de côté les modèles de langage les moins performants ;
- faire l'étude de la mise à jour à long terme ;
- augmenter le nombre de sources de données ;
- réviser l'algorithme de recherche de documents : plutôt que de se fier aux sources de nouvelles pour amener les bonnes nouvelles, faire d'abord des recherche sur le serveur du télédiffuseur pour connaître les sujets qui seront traités et ensuite faire les recherches concernant ces sujets.

En somme, les idées ne manquent pas pour améliorer ce système. Le plus grand enjeu est l'évaluation du système à long terme. Comme ce système est tout nouveau, il n'y a pas encore suffisamment de données disponibles pour réaliser une étude plus exhaustive. De plus, une étude utilisant une meilleure recherche de documents devra être effectuée mais elle n'est pas si facile à réaliser étant donné qu'une fois un bulletin enregistré, les tests portant sur celui-ci ne peuvent pas se faire avec ce qui est disponible sur Internet. En fait il faut préalablement télécharger les documents des sites de nouvelles. Évidemment, ce choix produit un biais important dans l'expérience.

RÉFÉRENCES

[Adda 1997] G. Adda, M. Adda-Decker, J.L. Gauvain, L. Lamel, « Text Normalization and Speech Recognition in French », Proceedings of Eurospeech, 1997, p 2711-2714

[Boulianne 2000] G.Boulianne, P. Dumouchel, J. Brousseau, P. Ouellet, « Le système de RAPT du CRIM », 12e Congrès francophone AFRIF-AFIA de Reconnaissance des Formes et Intelligence Artificielle (RFIA 2000), 2000

[Boulianne 2003] G. Boulianne, J.-F. Beaumont, P. Cardinal, M. Comeau, P. Ouellet, P. Dumouchel, « Automatic segmentation of film dialogues into phonemes and graphemes », Proceedings of Eurospeech, 2003, p 1241-1244

[Bourlard 1996] H. Bourlard, H. Hermansky, N. Morgan, « Toward increasing speech recognition error rates », Speech Communication, vol. 18, 1996, p 205-231

[Brousseau 2003] J. Brousseau, J.-F. Beaumont, G. Boulianne, P. Cardinal, C. Chapdelaine, M. Comeau, F. Osterrath, P.Ouellet, « Automated close-captioning of live TV broadcast news in French », Proceedings of Eurospeech, 2003, p 1245-1248

[CBC 2003] CBC Radio-Canada, « Jalons de l'histoire – les années 80 », http://cbc.radio-canada.ca/htmfr/historique/annees_80.htm, 2003

[Chen 1996] S. Chen, J. Goodman, « An Empirical Study of Smoothing Techniques for Language Modeling », Proceedings of the 34th Meeting of the Association for Computational Linguistics, ACL, 1996

[Chen 1998] S. F. Chen, K. Seymore, R. Rosenfeld, « Topic adaptation for language modeling using unnormalized exponential models », ICASSP, Seattle, USA, 1998, p 681-684

[Church 1988] K. Church, « A stochastic parts program and noun phrase parser for unrestricted text », Proceedings of the Second Conference on Applied Natural Language Procesing, 1988, p. 136-143

[Clarkson 1997] P. Clarkson, A. Robinson, « Language Model Adaptation Using Mixtures and an Exponentially Decaying Cache », Proceedings ICASSP, vol. 2, 1997, p 799-802

[Clarkson 1998] P.R. Clarkson, A.J. Robinson, « The Applicability of Adaptive Language Modelling for the Broadcast News Task », Proceedings ICSLP, 1998, p 1699-1702

[Cole 1996] R. Cole, et al., « Survey of the Sate of the Art in Human Language Technology », eds. http://cslu.cse.ogi.edu/HLTsurvey/HLTsurvey.html, Cambridge University Press, 1996

[Craven 2000] M. Craven, D. DiPasquo, D. Freitag, A. McCallum, T. Mitchell, K. Nigam, S. Slattery, « Learning to construct knowledge bases from the World Wide Web », Artificial Intelligence, vol. 118, 1999, p 69-113

[Descout 1992] Descout, R., Bergeron, R., Mérialdo, B., « Mediatex-Tasf : A closed captioning real-time service in French », ICSLP, 1992, p 1617-1620

[Federico 1996] M. Federico, « Bayesian Estimation Methods for N-gram Language Model Adaptation », Proceedings ICSLP, 1996, p 240-243

[Galescu 1998] L. Galescu, E. Ringer, J. Allen, « Rapid language model development for new task domains », Proceedings International Conference on Language Resources and Evaluation, 1998

[Gauvain 1995] J.L. Gauvain, L.F. Lamel and M. Adda-Decker, « Developments in Continuous Speech Dictation using the ARPA WSJ Task », Proceedings IEEE ICASSP, 1995, p 65-68

[Gildea 1999] D. Gildea, T. Hofmann, « Topic-Based Language Models Using EM », Eurospeech, vol. 5, 1999, p 2167-2170

[Good 1953] I.J. Good, « The Population Frequencies of Species and the Estimation of Population Parameters », Biometrika 40, 1953, p 237-264

[Iyer 1996] R. Iyer, M. Ostendorf, « Modeling Long Distance Dependence in Language : Topic Mixtures vs. Dynamic Cache Models », Proceedings ICSLP, vol. 1, 1996, p 236-239

[Janiszek 2000] D. Janiszek, F. Bechet, R. de Mori, « Integrating MAP and linear transformation for language model adaptation », Proceedings 6th International Conference on Spoken Language, vol. 2, 2000, p 895-898

[Jeffreys 1948] H. Jeffreys, « Theory of Probability », 2^{ième} édition, Oxford at the Clarendon Press, London, 1948

[Jelinek 1990] F. Jelinek, R. Mercer, S. Roukos, « Classifying Words for Improved Statistical Language Models », ICASSP, 1990, p 621-624

[Jelinek1997] F. Jelinek, « Statistical Methods for Speech Recognition », Cambridge : MIT Press, 1997

[Jin 1999] H. Jin, R. Schwartz, S. Sista, F. Walls, « Topic Tracking for Radio, TV Broadcast, and Newswire », Proc. Eurospeech, vol. 6, 1999, p 2439-2442

[Joachims 2002] T. Joachims, « Learning to Classify Text Using Support Vector Machines », Dissertation, Kluwer, 2002

[Jones 1981] K.S. Jones, « Information Retrieval Experiment », Butterworth and Co., 1981

[Jurafsky 1995] D. Jurafsky, C. Wooters, J. Segal, A. Stolcke, E. Fosler, G. Tajchman, N. Morgan, « Using a Stochastic Context-Free Grammar as a Language Model for Speech Recognition », Proceedings ICASSP, 1995, p 189-192

[Katz 1987] Katz, Slava M. « Estimation of probabilities from sparse data for the language model component of a speech recognizer », IEEE Transactions on Acoustics, Speech and Signal Processing, vol. 1, 1987, p 181-184

[Klakow 2000] D. Klakow, « Selecting articles from the language model training corpus », Proceedings ICASSP, 2000, p 1695-1698

[Kneser 1995] R. Kneser, H. Ney, « Improved backing-off for m-gram language modeling », Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, vol. 1, 1995, p 181-184

[Kneser 1997a] R. Kneser, J. Peters, D. Klakow, « Language Model Adaptation Using Dynamic Marginals », Eurospeech, 1997, p 1971-1974

[Kneser 1997b] R. Kneser, J. Peters, « Semantic Clustering for Adaptive Language Modeling », Proceedings ICASSP, 1997, p 779-782

[Knight 2001] S. Knight, G. Gorrell, M. Rayner, D. Milward, R. Koeling, I. Lewin, « Comparing grammar-based and robust approaches to speech understanding: a case study », Proceedings Eurospeech, 2001, p 1779-1782

[Kobayashi 1998] A. Kobayashi, K. Onoe, T. Imai, A. Ando, « Time Dependent Language Model for Broadcast News Transcription and its Post-Correction », ICSLP, 1998, p 2435-2438

[Kuhn 1990] R. Kuhn, R. de Mori, « A Cache-Based Natural Language Model for Speech Recognition », IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 12, 1990, p 570-583

[Lidstone 1920] G. Lidstone, « Note on the general case of the Bayes-Laplace formula for inductive or a posteriori probabilities », Transactions of the Faculty of Actuaries, vol. 8, 1920, p 182-192

[Lincoln 2003] C. Lincoln, « Notre souveraineté canadienne, Le deuxième siècle de la radiodiffusion canadienne », Comité du parlement canadien, 2003, chapitre 15, http://www.parl.gc.ca/InfoComDoc/37/2/HERI/Studies/Reports/herirp02/01a-cov2-f.htm

[Lippmann 1997] R. Lippmann, « Speech recognition by humans and machines », Speech Communication, vol. 22, 1997, p 1-15

[Mahajan 1999] M. Mahajan, D. Beefeman, D. Huang, « Improved topic-dependent language modeling using information retrieval techniques », Proceedings ICASSP, 1999, p 541-544

[Marti 2001] U. Marti, H. Bunke, « Using a statistical language model to improve the performance of an hmm-based cursive handwriting recognition system », International Journal of Pattern Recognition and Artificial Intelligence, vol. 15, 2001, p 65-90

[Neal 1998] R. Neal, G. Hinton, « A view of the EM algorithm that justifies incremental, sparse, and other variants », in M. I. Jordan, editor, Learning in Graphical Models, 1998, p 355-368

[Ney 1994] H. Ney, U. Essen, R. Kneser, « On structuring probabilistic dependences in stochastic language modeling », Computer, Speech, and Language, vol. 8, 1994, p 1-38

[Niesler 2002] T. Niesler, D. Willet, « Unsupervised Language Model Adaptation for Lecture Speech Transcription », ICSLP, 2002, p 1413-1416

[O'Shaughnessy 2003] D. O'Shaughnessy, « Interacting With Computers by Voice: Automatic Speech Recognition and Synthesis », Proceedings of the IEEE, vol. 91, 2003, p 1272-1305

[Pallett 1994] D.S. Pallett, J.G. Fiscus, W.M. Fisher, J.S. Garofolo, B.A. Lund, A,F, Martin, M.A. Przybocki, « 1994 Benchmark Tests for the ARPA Spoken Language Program », Proc. ARPA Spoken Language Systems Technology Workshop, 1995, p 5-35

[Rosenfeld 1996] R. Rosenfeld, « A Maximum Approach to Adaptive Statistical Language Modeling », Computer, Speech and Language, vol. 10, 1996, p 187-228

[Rosenfeld 2000] R. Rosenfeld, « Two decades of statistical language modeling : Where do we go from here? », Proceedings of the IEEE, vol. 88, 2000, p 1270-1278

[Salton 1991] G. Salton, « Developments in automatic text retrieval », Science, vol. 253, 1991, p 974-980

[Sawaf 2000] H. Sawaf, K. Schutz, H. Ney, « On the Use of Grammar Translation », Proc. of the 6th International Workshop on Parsing Technologies, 2000, p 231-241

[Seymore 1996a] K. Seymore, R. Rosenfeld, « Scalable Backoff Language Model », Proceedings of ICSLP, 1996, p 232-235

[Seymore 1996b] K. Seymore, R. Rosenfeld, « Scalable Trigram Backoff Language Models », Carnegie Mellon University Tech Report, 1996

[Seymore 1997] K. Seymore, R. Rosenfeld, « Using story topics for language model adaptation », Proceedings of Eurospeech, 1997, p 1987-1990

[Song 1999] F. Song, W. B. Croft, « A General Language Model for Information Retrieval », Research and Development in Information Retrieval, 1999, p 279-280

[Srihari 1993] R. K. Srihari, C. M. Baltus, « Incorporating Syntactic Constraints in Recognizing Handwritten Sentences », Proc. of the International Joint Conference on Artificial Intelligence, 1993, p 1262-1267

[Stolcke 1998] A. Stolcke, « Entropy-based Pruning Backoff Language Models », Proceedings of the ARPA Workshop on Human Language Technology, 1998, p 270-274

[Stolcke 2002] A. Stolcke, « SRILM – An Extensible Language Modeling Toolkit », Proceedings of International Conference of Spoken Language, 2002

[Vapnik 1995] V. N. Vapnik, « The Nature of Statistical Learning Theory », Springer, 1995.

[Walls 1999] F. Walls, H. Jin, S. Sista, R. Schwartz, « Topic Detection in Broadcast News », Eurospeech, vol. 6, 1999, p 2451-2454

[Witten 1991] I. H. Witten, T.C. Bell, « The zero-frequency problem: Estimating the probabilities of novel events in adaptive text compression », IEEE Transactions on Information Theory, vol. 37, 1991, p 1085-1094

[Young 1998] S. Young, L. Chase, « Speech recognition evaluation : A review of the U.S. CSR and LVCSR programmes », Computer Speech and Language, vol. 12, 1998, p 263-279

[Yu 2002] H. Yu, J. Han, K.C-C. Chang, « PEBL:Positive Example-Based Learning for Web Page Classification Using SVM », Proceedings ACM SIGKDD International Conference Knowledge Discovery in Databases, ACM Press, 2002, p 239-248

[Zhu 2001] X. Zhu, R. Rosenfeld, « Improving Trigram Language Modeling with the World Wide Web », Proceedings of ICASSP, 2001, p 592-597

ANNEXE A : SUJETS UTILISÉS POUR LA CLASSIFICATION INITIALE

Domaines	Sujets associés dans CEDROM-SNi
Le Monde	Accès à l'information et renseignements personnels Administration et finances publiques Aide internationale et humanitaire Ambassades, délégations et consulats Chefs d'États et de gouvernements Chefs de partis politiques Conditions et politiques économiques Conflits armés Corruption et abus de confiance Coups d'État Droits et libertés Défense nationale et armée Démographie et population Famille royale Frontières et territoires Géographie et géopolitique Immigrants, émigrants et réfugiés Mouvements et systèmes politiques Organisations internationales Terrorisme et assassinats politiques Torture et esclavage Traités et conventions Visites officielles Émeutes et manifestations
National	Accès à l'information et renseignements personnels Administration et finances publiques Agriculture et foresterie Agriculture et services connexes Aide internationale et humanitaire Aide sociale et assistés sociaux Autochtones et amérindiens Armes et armements Arrestations, opérations et brutalité Associations patronales et organismes d'affaires Chefs d'États et de gouvernements Chefs de partis politiques Chômage Conditions et politiques économiques Conditions sociales

	Constitution Corruption et abus de confiance Coûts de la santé Coûts des études et frais de scolarité Droits et libertés Défense nationale et armée Démographie et population Députés et représentants Fonction publique Frontières et territoires Géographie et géopolitique Hôpitaux psychiatriques et soins en santé mentale Hôpitaux, soins hospitaliers et urgences Immigrants, émigrants et réfugiés Industries forestières, du bois et des pâtes et papiers Industries manufacturières Industries militaires Infirmières et personnel de la santé Infrastructures publiques Langue et questions linguistiques Lois et règlements Ministères et ministres Minorités culturelles et linguistiques Nationalisme Organismes gouvernementaux Organismes sociaux, civiques et populaires Parlement, commissions et comités Partis politiques Patrimoine Personnalités politiques Politique et administration nationale Programmes politiques Programmes sociaux Protection civile et sécurité publique Présidents, administrateurs et conseils d'administration Pêcheries et chasse commerciale Relations autochtones/gouvernement Relations industrie/gouvernement Relations industrie/éducation Relations intergouvernementales Ressources naturelles et énergie Référendums Routes, autoroutes et ponts Secteurs publics et parapublics Service et travail social Services publics Services sociaux et de santé

	Sociétés d'état et entreprises publiques Soins à domicile Sondages et opinion publique Subventions et aide gouvernementale Syndicats Sénateurs et sénat Taxes et impôts Visites officielles Éducation Élections Émeutes et manifestations
Régional	Administration hospitalière Administration scolaire et administrateurs Aide sociale et assistés sociaux Architecture et urbanisme Arrestations, opérations et brutalité Associations patronales et organismes d'affaires Autochtones et amérindiens Chefs d'États et de gouvernements Chefs de partis politiques Chômage Conditions et politiques économiques Conditions sociales Conseils municipaux Constitution Corruption et abus de confiance Coûts de la santé Coûts des études et frais de scolarité Démographie et population Députés et représentants Fonction publique Fusions municipales Hôpitaux psychiatriques et soins en santé mentale Hôpitaux, soins hospitaliers et urgences Infirmières et personnel de la santé Infrastructures publiques Langue et questions linguistiques Logement Lois et règlements Maires et élus locaux Ministères et ministres Minorités culturelles et linguistiques Nationalisme Organismes gouvernementaux Organismes sociaux, civiques et populaires Parlement, commissions et comités Partis politiques

	Patrimoine Personnalités politiques Politique et administration locale et municipale Politique et administration régionale Programmes politiques Programmes sociaux Protection civile et sécurité publique Présidents, administrateurs et conseils d'administration Régionalisme et décentralisation Relations autochtones/gouvernement Relations industrie/gouvernement Relations industrie/éducation Relations intergouvernementales Ressources naturelles et énergie Référendums Routes, autoroutes et ponts Secteurs publics et parapublics Service et travail social Services publics Services sociaux et de santé Sociétés d'état et entreprises publiques Soins à domicile Sondages et opinion publique Subventions et aide gouvernementale Syndicats Taxes et impôts Villes et quartiers Éducation Élections Émeutes et manifestations
Culture	Arts et culture Arts visuels Cinéma Concours Cuisine et restaurants Danse Festivals Fêtes et festivités Habitat, jardinage et décoration Humour Hébergement, restauration et industrie touristique Industrie de la musique Industrie de la production télévisuelle Industrie du cinéma Littérature et livres Mode et beauté Musique

	Médias et information Parcs et espaces verts Radio et télévision Services de divertissements et de loisirs Théâtre Variétés et événements culturels Villes et quartiers Voyages, aventures et tourisme
Économie	Agences de recouvrement Banque centrale, banques et services bancaires Bourse et marché des changes Chômage Colloques, congrès et réunions d'affaires Commerce de détail Commerce extérieur Commerce électronique Comptabilité et fiscalité Concurrence Conditions et politiques économiques Conglomérats et holdings Coopératives Droit commercial et des affaires Entreprises familiales Entreprises privées Exploration minière et mines Exportations et importations Fermetures d'entreprises Finances des entreprises Finances et placements personnels Fraudes et crimes économiques Gens d'affaires et entrepreneurship Gestion des ressources humaines Gestion et administration des affaires Grèves et manifestations étudiantes Industrie de la musique Industrie de la production télévisuelle Industrie du cinéma Industries aéronautiques et aérospatiales Industries chimiques Industries culturelles Industries de l'alimentation et des boissons alcoolisées et gazeuses Industries de l'automobile Industries de l'environnement Industries de l'informatique et de l'électronique Industries de l'énergie Industries de transformation des métaux Industries des articles de sport

	Industries des produits de caoutchouc Industries des produits de matière plastique Industries des produits métalliques Industries des véhicules récréatifs Industries du cuir et de la chaussure Industries du meuble Industries du tabac Industries forestières, du bois et des pâtes et papiers Industries manufacturières Industries militaires Industries pharmaceutiques et biotechnologiques Industries pétrolières et pétrochimiques Industries textiles et du vêtement Inflation, prix et salaires Investissements Mises à pied et congédiements PIB/PNB PME Partenariats et alliances stratégiques Privatisations Productivité Relations de travail Salaires Services et produits financiers Taux d'intérêt Taxes et impôts Théories et systèmes économiques Travail autonome Travail et emploi Travail à domicile et télétravail Travailleurs et conditions de travail Économie et gestion Économie locale Économie mondiale Économie nationale Économie régionale
Sports	Accidents et sécurité dans les sports Industries des articles de sport Sports et loisirs
Météo	Accidents, catastrophes et climat Climat et catastrophes naturelles Lacs, cours d'eau et océans Milieux protégés Météo Parcs et espaces verts Pluie et verglas

	Tempêtes de neige et déneigement Tremblements de terre Vents violents et tornades Volcans

ANNEXE B : FORMAT ARPA DES MODÈLES *N*-GRAMMES AVEC REPLI

Ceci est le format standard utilisé par plusieurs systèmes pour représenter un modèle de langage à base de *n*-grammes. Ce standard fut conçu par Doug Paul alors au laboratoire Lincoln du MIT conduisant des recherches financées par l'Advanced Research Project Agency (ARPA) du département de la défense américaine.

Comme il n'est normalement pas possible (étant donné le grand nombre de *n*-grammes) de générer une probabilité pour tous les *n*-grammes possibles, une stratégie basée sur la technique de repli est utilisée. Pour calculer la probabilité d'un *n*-gramme manquant, la probabilité du repli est utilisée en combinaison avec la probabilité du *n*-gramme d'ordre inférieur.

Le format est le suivant :

```
\data\
ngram 1=n1
ngram 2=n2
…
ngram N=nN
\1-grams:
log10(p) w1 [log10(bow)]
…
\2-grams:
log10(p) w1 w2 [log10(bow)]
…
\N-grams:
log10(p) w1 … wN
…
\end\
```