Exploration et utilisation d'informations distantes dans les modèles de langage statistiques

Armelle Brun, David Langlois, Kamel Smaïli

Université Nancy 2, LORIA
{brun ; langlois ; smaili}@loria.fr

Résumé

Dans le cadre de la modélisation statistique du langage, nous montrons qu'il est possible d'utiliser un modèle n-grammes avec un historique qui n'est pas nécessairement celui avec lequel il a été appris. Par exemple, un adverbe présent dans l'historique peut ne pas avoir d'importance pour la prédiction, et devrait donc être ignoré en décalant l'historique utilisé pour la prédiction. Notre étude porte sur les modèles n-grammes classiques et les modèles n-grammes distants et est appliquée au cas des bigrammes. Nous présentons quatre cas d'utilisation pour deux modèles bigrammes : distants et non distants. Nous montrons que la combinaison linéaire dépendante de l'historique de ces quatre cas permet d'améliorer de 14 % la perplexité du modèle bigrammes classique. Par ailleurs, nous nous intéressons à quelques cas de combinaison qui permettent de mettre en valeur les historiques pour lesquels les modèles que nous proposons sont performants.

Mots-clés : modélisation statistique du langage, modèles distants, combinaison linéaire.

Abstract

In the framework of statistical language modeling, we show that it is possible to use n-gram models with a history different to the one used during training. Our study deals with classical and distant n-gram models and is restricted to bigram models. We present four use cases for two bigram models : distant and non distant. By using the linear combination, we show an improvement of 14 % in terms of perplexity compared to the classic bigram model. Moreover, a study has been performed in order to emphasize the histories for which our models are efficient.

Keywords: statistical language modeling, distant models, linear combination.

1. Introduction

Un modèle statistique de langage définit la vraisemblance $P(W) = P(w_1, \ldots, w_N)$ d'une suite de mots $W = w_1, \ldots, w_N$, issus d'un lexique. Cette vraisemblance est évaluée par un produit de probabilités $P(w_i \mid h_i)$, où w_i est le i^{e} mot de la suite et h_i la suite de mots précédant w_i[1]. C'est la distribution de probabilités P qui définit un modèle statistique de langage. Les plus connus de ces modèles sont les modèles dits n-grammes qui réduisent l'historique à ses $n-1$ derniers mots afin d'apporter une réponse au manque de données (Jelinek et Mercer, 1980).

$$P(W) = \prod_{i=1}^{N} P(w_i|h_i) = \prod_{i=1}^{N} P(w_i|w_{i-n+1} \ldots w_{i-1}) \quad (1)$$

Un modèle n-grammes est performant du fait qu'il prend en compte les contraintes syntagma-

[1] Désormais, h_i sera appelé *historique*.

tiques locales qui sont souvent très fortes (comme l'accord entre un article et un nom, un pronon personnel sujet et un verbe, etc.). Cependant, ces modèles restent limités en raison de la taille de l'historique, un effet d'horizon empêche donc de capter les contraintes allant au delà des $n-1$ mots précédents. L'allongement de l'historique est une solution qui ne peut être systématisée du fait du manque de données (Martin *et al.*, 1997). Un deuxième problème est le manque de généralisation. En effet, si un modèle trigrammes nous donne des informations sur les suites possibles de « je la », il n'est pas possible d'utiliser ces informations pour prédire les suites possibles de « je le ». Les modèles n-classes (Smaïli *et al.*, 1999) sont une réponse possible à ce problème à condition que le passage au niveau des classes n'induise pas une trop grande perte d'information.

Nous abordons dans cet article un autre problème, non explicitement décrit dans la littérature à notre connaissance : pour une valeur de n donnée, le modèle utilise toujours systématiquement la même partie de l'historique. Par exemple, dans l'historique « le chien furieux », un modèle trigrammes utilisera « chien furieux »pour prédire la suite. Or, un modèle trigrammes nous apprend aussi que « le chien »est souvent suivi (par exemple) par « aboie », « grogne », « attaque ». Il serait donc judicieux de faire abstraction de og furieux »et d'utiliser og le chien »pour la prédiction[2]. Mais ce choix est très dépendant de l'historique. En effet, si l'historique og chien furieux »est fortement représenté dans le corpus, il sera plus judicieux d'utiliser cet historique pour la prédiction. Une autre solution consiste bien sûr à ne pas faire de choix et à utiliser les deux possibilités.

Dans le même ordre d'idées, les modèles n-grammes distants (Huang *et al.*, 1993), que nous détaillerons section 2, utilisent une sous partie de l'historique constituée de $n-1$ mots situés d mots en arrière. Ces modèles ont les mêmes défauts que les modèles n-grammes avec le problème supplémentaire que la distance relâche les contraintes locales. De fait, ils sont moins performants que les modèles n-grammes classiques (non distants), comme le confirmera cet article. Pourtant, il est ici aussi possible d'utiliser le même raisonnement que précédemment. En effet, dans le cadre des trigrammes, l'historique og le chien »est souvent suivi par un verbe. Mais il peut aussi survenir un adjectif. On aura ainsi de nombreux historiques « le chien furieux », « le chien fou », « le chien bavant », etc. qui seront suivis par un verbe. Un modèle trigrammes distant de distance 1 peut renseigner sur cette situation. Donc, quand on rencontre l'historique « le chien », peut-être est-il judicieux d'utiliser le modèle trigrammes distant pour prédire, peut-être un verbe.

L'idée centrale de cet article est d'utiliser un modèle de langage avec un historique qui n'est pas nécessairement celui avec lequel il a été appris. Ainsi la probabilité d'un mot w_i sachant son historique est estimée selon les différents cas présentés dans le tableau 1. À titre d'exemple, le cas 2 correspond à un modèle n-grammes classique utilisant un historique distant (noté h_i^{-1}). Ainsi, en reprenant l'exemple précédent « le chien furieux aboie », ce cas correspond à la prédiction du mot « aboie »via le modèle trigrammes classique mais en s'appuyant sur la suite « le chien ». Nous combinerons ces quatres modèles par la biais d'une combinaison linéaire dépendante de l'historique.

Dans la section 2, nous rappelons le formalisme général des modèles n-grammes distants. Dans la section suivante nous formalisons l'utilisation d'un modèle n-grammes distant ou non avec un historique qui correspond ou non à celui utilisé dans l'apprentissage. Dans la section 4, nous nous intéressons à la combinaison de ces modèles. Ensuite, nous présentons les données support

[2] Bien sûr, un modèle 4-grammes permettrait peut-être de résoudre le problème, sauf si le modèle ne propose aucune information supplémentaire sur l'historique og le chien furieux »par rapport au modèle trigrammes.

Historique / Modèle	h_i classique	h_i^{-1} distant
n-grammes	cas 1 : $P(w_i \mid w_{i-n+1} \dots w_{i-1})$	cas 2 : $P(w_i \mid w_{i-n} \dots w_{i-2})$
n-grammes distant[a]	cas 3 : $P_d(w_i \mid w_{i-n+1} \dots w_{i-1})$	cas 4 : $P_d(w_i \mid w_{i-n} \dots w_{i-2})$

[a] Voir la section 2 pour la formalisation du modèle n-grammes distant P_d

Tableau 1. Présentation des quatre cas d'utilisation des modèles n-grammes

de nos travaux. Puis, nous reportons et analysons les résultats que nous avons obtenus. La section 6 analyse quelques cas de combinaison de modèles de langage pour lesquels la combinaison a un fort intérêt. Enfin, nous concluons cette étude et présentons quelques perspectives.

2. Les modèles distants

Nous présentons d'abord la formalisation des modèles n-grammes distants. Ces derniers permettent de modéliser une relation distante valuée entre mots ou groupes de mots. Un modèle n-grammes distant de distance d modélise le lien statistique entre un historique de $n-1$ mots séparé du mot à prédire par exactement d mots (dans un modèle n-grammes classique cette séparation est de $d = 0$ mots). Les probabilités P_d, tout comme pour un modèle n-grammes, sont estimées sur un corpus d'apprentissage :

$$P_d(w_i \mid w_{i-n+1-d} \dots w_{i-1-d}) = \frac{N_d(w_{i-n+1-d} \dots w_{i-1-d}, w_i)}{N(w_{i-n+1-d} \dots w_{i-1-d})} \quad (2)$$

Où $N_d(w_{i-n+1-d} \dots w_{i-1-d}, w_i)$ correspond au nombre de fois où le mot w_i suit la suite de mots $w_{i-n+1-d} \dots w_{i-1-d}$ dans le corpus d'apprentissage, et d est la distance séparant w_i de son historique[3]. Comme nous l'avons mentionné précédemment, ces modèles utilisés seuls sont bien évidemment moins performants que les modèles n-grammes classiques (de distance 0) du fait qu'ils n'exploitent pas du tout la relation entre le mot et son contexte gauche immédiat (Langlois, 2002).

Nous n'aborderons pas dans ce travail d'autres modèles utilisant des parties distantes de l'historique (modèles triggers, cache) car ces modèles perdent une grande partie des relations syntagmatiques locales entre les mots au profit d'une relation plus « sémantique ». Nous reportons le lecteur à (Federico et De Mori, 1997) pour une description plus détaillée de ces modèles.

3. Formalisation des modèles

Comme nous l'avons présenté dans l'introduction nous proposons dans cet article d'utiliser des modèles n-grammes distants de distance $m = d$ ($d \geq 0$, m pour *m*odèle) avec une distance d'utilisation en test $u = \delta$ ($\delta \geq 0$, u pour *u*tilisation) différente de celle du modèle.

Nous formalisons les 4 possibilités évoquées dans l'introduction en généralisant l'équation 2 :

$$P_{m=d\_u=\delta}(w_i|h_i) = P_d(w_i|w_{i-n+1-\delta} \cdots w_{i-1-\delta}) \quad (3)$$

Le modèle défini par $P_{m=d\_u=\delta}$ sera noté par $M_{m=d\_u=\delta}$. Suivant ces notations, les 4 modèles évoqués sont :

[3] Nous ne présentons pas ici les différentes méthodes utilisées pour estimer la probabilité n-grammes d'un mot quand le n-gramme n'est pas présent, ou trop peu présent dans le corpus d'apprentissage (Chen et Goodman, 1999).

– $M_{m=0\_u=0}$: le modèle n-grammes classique, dans son utilisation standard, non distante. Sa notation sera simplifiée en M ;
– $M_{m=0\_u=1}$: le modèle n-grammes classique, dans son utilisation distante ;
– $M_{m=1\_u=1}$: le modèle n-grammes distant dans son utilisation standard. Sa notation sera simplifiée en M_1 ;
– $M_{m=1\_u=0}$: le modèle n-grammes distant dans son utilisation non distante.

4. Combinaison des modèles de langage

Dans la littérature, plusieurs méthodes de combinaison de modèles de langage sont proposées. Nous pouvons par exemple citer le maximum d'entropie (Rosenfeld, 1994) ou encore la combinaison linéaire (Jelinek *et al*., 1991). Dans ce qui suit, nous choisissons d'utiliser la combinaison linéaire. En effet, celle-ci nous permettra facilement d'analyser *a posteriori* le poids associé à chacun des modèles, ce que nous ne pourrions pas faire en utilisant le maximum d'entropie.

La combinaison linéaire associe un ou plusieurs poids à chaque modèle. Dans cet article, nous étudierons deux cas de combinaison linéaire :
– la combinaison linéaire non dépendante de l'historique qui associe un seul poids à chaque modèle (tout historique confondu). Les expérimentations concernant la combinaison linéaire indépendante de l'historique seront données à titre de comparaison avec la combinaison dépendante de l'historique (voir point suivant), afin de valider le fait que l'utilisation des 4 modèles est fortement dépendante de l'historique ;
– la combinaison linéaire dépendante de l'historique (Jelinek et Mercer, 1980), qui associe à chaque historique distinct, un poids pour chaque modèle. Cette combinaison permet d'accorder plus ou moins d'importance aux modèles, en fonction de l'historique. Dans ce cas, la probabilité d'un mot sachant son historique, appliquée à notre problème, est calculée de la façon suivante :

$$P(w_i \mid h_i) = \sum_{d\in\{0,1\}} \sum_{\delta\in\{0,1\}} \lambda_{h_i,m=d\_u=\delta} P_{m=d\_u=\delta}(w_i|h_i) \qquad (4)$$

où les $\lambda_{h_i,m=d\_u=\delta}$ sont tels que $0 \leq \lambda_{h_i,m=d\_u=\delta} \leq 1$, somment à 1 et sont définis pour chaque historique h_i. Ces poids sont optimisés sur un corpus de développement en utilisant l'algorithme *Expectation-Maximisation* (Dempster *et al*., 1977) (dans nos travaux, nous avons utilisé l'outil `interpolate` de Carnegie Mellon University (Clarkson et Rosenfeld, 1997)).

On peut légitimement se poser la question de la sous-représentation des données dans le corpus de développement pour le calcul des poids optimaux (un poids par modèle et par historique). Nous verrons par la suite que la masse de données que nous utilisons en développement nous permet tout de même d'associer un jeu de poids distincts pour les historiques fréquents.

5. Données et résultats

Les données utilisées pour l'apprentissage et l'évaluation des quatre modèles sont issues du journal *Le Monde*. Pour le corpus d'apprentissage, 12 années sont utilisées (de 1987 à 1998), ce qui correspond à un peu plus de 288 M mots. Le corpus de développement correspond aux années 1999-2001 (79 M mots). Le corpus de test, quant à lui, est l'année 2002 (27 M mots). Le vocabulaire est composé des 65 K mots utilisés dans le cadre de la campagne ESTER. Le choix

Modèle	Perplexité sur le corpus de test
M	164.7
M_1	499.4
$M_{m=1\_u=0}$	2403.8
$M_{m=0\_u=1}$	20632.8

Tableau 2. Perplexité de chacun des modèles de langage

de ce vocabulaire a été fait dans l'optique d'une utilisation à venir des modèles décrits ici dans le système de transcription automatique ANTS (Fohr *et al.*, 2004).

Dans le cadre de ces premières expérimentations, nous testons l'idée avec des modèles bigrammes. Ce choix a été dicté par le souci de pouvoir plus facilement analyser les historiques en fonction des poids respectifs associés aux modèles : un historique de 1 mot est plus simple à analyser pour un humain. Pour tous les modèles, nous avons utilisé la méthode de discounting *Absolute Discounting* (Ney et Essen, 1991).

5.1. Évaluation des quatre modèles de langage

Nous évaluons les quatre modèles de langage en exploitant la mesure de perplexité (Jelinek et Mercer, 1980). Les valeurs de perplexité de chacun de ces modèles, sur le corpus de test, sont présentées dans le tableau 2.

La perplexité de référence que nous obtenons est de 164.7, qui correspond à celle du modèle bigramme classique. Les trois autres perplexités obtenues sont largement supérieures à celle-ci, ce qui confirme la littérature pour le modèle M_1 (Langlois, 2002). Le modèle qui obtient la seconde meilleure perplexité est le modèle distant classique. Ce résultat semble normal, puisque dans ce modèle des relations distantes sont modélisées et ensuite exploitées telles quelles. Enfin, les deux modèles que nous proposons ont des perplexités beaucoup plus élevées. Cependant le modèle $M_{m=1\_u=0}$ a une perplexité dix fois plus faible que le modèle $M_{m=0\_u=1}$. Il est difficile de donner une explication à cet écart. Nous pouvons simplement avancer que, dans les deux types d'exemples donnés en introduction, l'un est globalement plus pertinent que l'autre.

Nous évaluons maintenant les performances de la combinaison linéaire d'abord indépendante de l'historique, puis dépendante de l'historique.

5.2. Combinaison linéaire indépendante de l'historique

Dans un premier temps nous évaluons le gain en perplexité en combinant linéairement les modèles sans tenir compte de l'historique. Un seul poids est utilisé pour chacun des modèles. Nous prenons en compte dans cette combinaison tous les modèles, même le plus faible, car nous cherchons à savoir pour quels historiques chacun des modèles peut être utile. Les valeurs de perplexité obtenues sont présentées dans le tableau 3.

Nous pouvons remarquer que l'amélioration de la perplexité dans le cas d'un poids fixe pour chacun des modèles permet d'améliorer la perplexité de 4.6 % par rapport au modèle bigrammes classique M. Il est donc clair que la combinaison simple de ces modèles n'a pas beaucoup d'impact. D'ailleurs, le poids du modèle de base M reste élevé (0.9). Par ailleurs, l'introduction des deux modèles $M_{m=0\_u=1}$ et $M_{m=1\_u=0}$ n'a globalement aucun impact. Les poids de ces modèles dans la combinaison des 4 modèles est négligeable.

Modèles utilisés	Poids des modèles sur le corpus de développement				Perplexité sur le corpus de test
	M	$M_{m=0\_u=1}$	$M_{m=1\_u=0}$	M_1	
$M + M_{m=1\_u=0}$	0.991	–	0.009	–	164.7
$M + M_{m=0\_u=1}$	0.986	0.014	–	–	164.4
$M + M_1$	0.870	–	–	0.130	157.2
$M + M_{m=0\_u=1} +$ $M_{m=1\_u=0} + M_1$	0.878	0.002	0.003	0.117	157.2

Tableau 3. Perplexité correspondant à la combinaison linéaire de modèles

Modèles utilisés	Nombre d'historiques fréquents		
$M + M_{m=0\_u=1}+$	*100K*	*200K*	*300K*
	144.1	142.9	142.3
	400K	*500K*	*600K*
$M_{m=1\_u=0} + M_1$	141.9	141.6	141.4
	700K	*800K*	*900K*
	141.3	141.2	141.1

Modèles utilisés	Nombre d'historiques fréquents = 900K
$M + M_{m=1\_u=0}$	163.7
$M + M_{m=0\_u=1}$	159.8
$M + M_1$	143.3

Tableau 4. Perplexité correspondant à la combinaison linéaire de modèles dépendante de l'historique

5.3. Combinaison linéaire dépendante de l'historique

Pour la combinaison linéaire dépendante de l'historique, on ne peut raisonnablement pas estimer un jeu de poids pour chaque historique (plusieurs milliards dans notre cas). La littérature propose de classer les historiques en fonction de leur fréquence (Jelinek et Mercer, 1980). Mais ceci ne convient pas à notre objectif car cette classification rassemble des historiques qui peuvent être très différents sur le plan syntaxique ou lexical. Or, nous cherchons à déterminer l'importance du rôle de chaque modèle pour un historique donné. Nous proposons donc de créer une classe pour chaque historique fréquent, et de créer une dernière classe pour les autres historiques, considérés comme peu fréquents. Nous étudions l'influence du nombre de classes d'historiques sur la perplexité.

Nous présentons dans le tableau 4 les valeurs de perplexité obtenues lorsque le poids de chacun des modèles est dépendant de l'historique du mot à prédire et en fonction du nombre de classes étudiées (de 100K à 900 K). Les poids des différents modèles, comme précédemment, ont été optimisés sur le corpus de développement, en utilisant l'algorithme EM.

L'amélioration de la perplexité de 14.3 % (avec 900K classes d'historiques), montre la grande influence sur les performances de la prise en compte de l'historique pour l'estimation des poids. Nous notons toutefois que utiliser une classe pour chaque historique fréquent montre ses limites. En effet, rapidement, l'accroissement du nombre de classes n'a plus d'impact sur la perplexité. Or, pour 900K historiques ainsi pris en compte, il reste potentiellement plus de 4 milliards d'historiques non différenciés par les poids (tous contenus dans une seule classe). Nous nous intéresserons à la classification de ces historiques dans nos travaux futurs. Nous proposons dans la section suivante une première étude manuelle des historiques en fonction des poids associés aux modèles.

Exemple	Historique	Mots suivants	Exemple	Historique	Mots suivants
1	et donc	à, le, la l', de, et	4	à ne	pas, plus rien, jamais
2	au coeur	du, de_la, de_l' de, d', des, même	5	sa propre	liste, vie monnaie, armée
3	du commerce	extérieur, mondial de, international	6	du territoire	national, qui, en français, des

Tableau 5. Étude des poids associés à certaines historiques

6. Étude de quelques combinaisons de modèles

Nous nous penchons maintenant sur l'étude de quelques poids obtenus pour deux combinaisons, la première combine le modèle de base M et le modèle $M_{m=1\_u=0}$, la seconde combinant le modèle M et le modèle $M_{m=0\_u=1}$.

6.1. Combinaison des modèles M et $M_{m=1\_u=0}$

Dans le cadre de cette étude, nous combinons le modèle de langage bigrammes classique et un modèle bigrammes distant de distance 1 que nous exploitons avec un historique non distant ($M_{m=1\_u=0}$).

Après optimisation sur le corpus de développement, 78.2 % des historiques conservent un poids ≥ 0.9 au modèle bigramme classique, 1.2 % des historiques se voient affecter un poids de 1 au modèle $M_{m=1\_u=0}$ (donc aucun poids au modèle bigramme classique) et 9.6 % des historiques accordent un poids non négligeable au modèle $M_{m=1\_u=0}$ (poids supérieur à 0.3).

Nous allons tout d'abord nous intéresser aux historiques pour lesquels les deux modèles ont des poids non nuls : chacun des deux étant pris en compte pour prédire un mot. C'est le cas par exemple de l'historique `et donc`. Le modèle bigramme classique M et le modèle $M_{m=1\_u=0}$ sont utilisés avec un poids de 0.65 et 0.35 respectivement. Les mots suivant le plus couramment `et donc` dans le corpus de développement sont présentés dans l'exemple 1 du tableau 6.1, ils sont à la fois très courants quand `donc` est en contexte gauche direct (modèle M) dans le corpus d'apprentissage mais également très courants quand `donc` est à une distance de 1 (modèle $M_{m=1\_u=0}$). Les poids obtenus nous paraissent donc justifiés.

Nous nous penchons maintenant sur les cas où un poids nul est affecté au modèle $M_{m=1\_u=0}$ et un poids de 1 au modèle bigramme classique M. C'est le cas de l'historique `au coeur`. Ce contexte est la plupart du temps suivi d'un des mots de l'exemple 2 du tableau 6.1. Il est alors évident qu'un modèle distant est inutile pour prédire le mot qui suit directement `coeur`. Nous pouvons également citer le cas de l'historique `du commerce` (exemple 3). Il est alors évident que dans ce contexte, un modèle distant pour prédire le mot suivant directement `commerce` n'a aucun intérêt car la relation contigüe entre `commerce` et la suite est très forte.

Intéressons nous maintenant au cas opposé de ce dernier : le modèle de langage bigramme classique s'est vu affecter un poids nul, la totalité du poids revenant au modèle $M_{m=1\_u=0}$. Un exemple que nous pouvons montrer est l'historique `à ne` où le modèle bigramme distant, et uniquement lui, est utilisé (voir exemple 4). Les mots suivant `ne` à une distance 0 sont exactement les mêmes mots qui le suivent à une distance 1 dans le corpus d'apprentissage, dans des suites comme : *ne sont pas, ne peut pas, ne peut plus*.

6.2. Combinaison des modèles M et $M_{m=0\_u=1}$

Nous nous intéressons maintenant à la combinaison du modèle bigramme classique M et du modèle bigramme classique utilisant un historique distant ($M_{m=0\_u=1}$).

Dans le cas de cette combinaison, 77.4 % des historiques conservent un poids ≥ 0.9 au modèle bigramme classique. 2.3 % des historiques se voient affecter un poids de 1 au modèle $M_{m=0\_u=1}$ et 13.3 % des historiques accordent un poids non négligeable au modèle $M_{m=0\_u=1}$ (poids supérieur à 0.3).

Le premier cas auquel nous nous intéressons est le cas classique pour lequel à la fois le bigramme classique et le modèle $M_{m=0\_u=1}$ ont chacun un poids non nul. Le premier exemple que nous montrons respecte la structuration « *article adjectif nom_commun* ». C'est le cas de l'historique `sa propre`, voir exemple 5. Un article est souvent suivi directement d'un nom commun. Il en est de même pour un adjectif. Il est donc logique que le modèle non distant (M et $M_{m=0\_u=1}$) utilise ces deux informations pour conforter sa prédiction (d'un nom commun) et associe aux deux modèles des poids respectifs de 0.61 et 0.39.
Un cas similaire que nous évoquons est l'historique `nous nous`. Dans ce cas, il est logique que les deux modèles aient des poids très proches, puisque dans ce cas ce sont exactement les mêmes historiques qui sont utilisés (le poids de chacun des deux modèles est de 0.5).

Le second cas que nous présentons est celui pour lequel le modèle bigramme classique a un poids de 1 et le modèle $M_{m=0\_u=1}$ a un poids nul. C'est le cas pour l'historique `le professeur` dont le mot à suivre sera prédit uniquement en utilisant le mot `professeur`. En effet, cet historique est dans la majorité des cas suivi d'un nom propre. Étant donné que le mot `le` n'est jamais suivi d'un nom propre, ce modèle se voit affecter un poids nul.
Nous pouvons citer un autre cas, l'historique `du territoire`, voir exemple 6, pour lesquels un bigramme classique exploitant l'historique `du` est bien évidemment inutile.

Enfin, nous nous intéressons au cas inverse du précédent, le modèle bigramme classique a un poids nul et le modèle $M_{m=0\_u=1}$ a quant à lui un poids de 1. Un exemple est l'historique `y en`. Cet historique est suivi des mots `a, avait, aura, eut, ait`. Il est évident que dans ce cas, le bigramme classique (utilisant uniquement le mot `en`) ne peut prédire efficacement les mots à venir. Cependant, le mot `y` le permet. Un autre exemple intéressant concerne les historiques où le second mot de l'historique est le mot inconnu. Dans ces cas, le modèle bigramme classique n'a aucun poids, l'utilisation du mot inconnu comme historique n'apporte pas assez d'information et dans ce cas, la prédiction distante a un poids de 1.

7. Conclusion

Dans cet article, nous avons proposé d'exploiter des modèles de langage n-grammes en fondant la prédiction sur des historiques de distance potentiellement différente de celle de l'apprentissage. L'originalité de cet article est l'utilisation de modèles de langage pour lesquels la distance d'apprentissage est différente de la distance d'utilisation. Nous avons montré que bien évidemment de tels modèles, utilisés seuls, ont une perplexité moins bonne que des modèles de langage non distants et distants classiques. Cependant, lorsque ces derniers sont utilisés en combinaison avec un modèle de langage non distant et/ou un modèle de langage distant classiques, des améliorations de la perplexité sont obtenues. Lorsque cette combinaison exploite un seul poids par modèle (non dépendant de l'historique), l'amélioration de la perplexité est de 4.6 %. Lorsque la combinaison utilise un poids par modèle et par historique distinct, l'amélioration

de la perplexité dépasse 14 %.

Nous avons ensuite étudié quelques poids de combinaison des modèles et analysé les historiques auxquels ils correspondaient. Cette étude nous a permis de nous conforter dans l'idée que l'utilisation des modèles que nous proposons permet d'améliorer les performances des modèles de langage dans des cas bien précis d'historique où un modèle de langage classique est moins performant. Nos travaux vont maintenant s'étendre à l'étude de modèle n-grammes d'ordre supérieur à celui des modèles bigrammes, afin d'étudier l'apport de la distance dans de tels modèles. Des travaux sur les modèles trigrammes ont déjà été menés, nous montrant des améliorations proportionnellement similaires à celles présentées dans cet article. Après étude de ces modèles, et amélioration du regroupement des historiques, nous pourrons intégrer ces derniers dans un système de reconnaissance de la parole afin d'évaluer leur apport dans un tel système.

Références

CHEN S. F. et GOODMAN J. (1999). « An empirical study of smoothing techniques for language modeling ». In *Computer Speech and Language*, 13, 359–394.

CLARKSON P. R. et ROSENFELD R. (1997). « Statistical language modeling using the CMU-Cambridge toolkit ». In *Proceedings of the European Conference on Speech Communication and Technology*, volume 5 : ESCA. ESCA, Rhodes, Greece, p. 2707–2710.

DEMPSTER P., LAIRD N. et RUBIN D. (1977). « Maximum Likelihood from Incomplete Data via the EM Algorithm ». In *Journal of Royal Statistical Society*, 39, 1-38.

FEDERICO M. et DE MORI R. (1997). *Spoken dialogues with computers*, chapter Language Modelling, p. 199–230. Academic Press.

FOHR D., MELLA O., ILLINA I. et CERISARA C. (2004). « Experiments on the accuracy of phone models and liaison processing in a French broadcast news transcription system ». In *8th International Conference on Spoken Language Processing - ICSLP'2004, Jeju, South Corea*.

HUANG X., ALLEVA F., HON H., HWANG M., LEE K. et ROSENFELD R. (1993). « The SPHINX-II Speech Recognition System : An Overview ». In *Computer, Speech and Language*, 2, 137-148.

JELINEK F. et MERCER R. (1980). « Interpolated estimation of markov source parameters from sparse data ». In *Pattern Recognition in Practice*, p. 381-397.

JELINEK F., MERIALDO S., ROUKOS S. et STRAUSS M. (1991). « A dynamic language model for speech recognition ». In *Proceedings of the 4th DARPA Speech and Natural Language Workshop*. Pacific Grive, California, p. 293-295.

LANGLOIS D. (2002). *Notions d'événements distants et d'événements impossibles en modélisation stochastique du langage : application aux modèles n-grammes de mots et de séquences*. PhD thesis, Université Henri Poincaré - NancyI.

MARTIN S. C., LIERMANN J. et NEY H. (1997). « Adaptative topic-dependent language modelling using word-based varigrams ». In *Proceedings of the European Conference on Speech Communication and Technology*, volume 3. Rhodes, Grèce, p. 1447–1450.

NEY H. et ESSEN U. (1991). « On smoothing techniques for bigram-based natural language processing ». In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, volume 2. Toronto, p. 825–828.

ROSENFELD R. (1994). *Adaptative Statistical Language Modeling : A Maximum Entropy Approach*. PhD thesis, School of Computer Science, Carnegie Mellon University.

SMAÏLI K., BRUN A., ZITOUNI I. et HATON J.-P. (1999). « Automatic and manual cluster-

ing for large vocabulary speech recognition : a comparative study ». In *Proceedings of the European Conference on Speech Communication and Technology*, volume 4. Budapest, p. 1795–1798.