Apport d'un modèle de langage statistique pour la reconnaissance de l'écriture manuscrite en ligne

Freddy Perraud (1), Emmanuel Morin (2),
Christian Viard-Gaudin (3) et Pierre-Michel Lallican (1)

(1) Société Vision Objects - 9, rue du Pavillon - 44980 Sainte Luce sur Loire
{freddy.perraud, pmlallican}@visionobjects.com

(2) Institut de Recherche en Informatique de Nantes
2, rue de la Houssinière - BP 92208 - 44322 Nantes Cedex 3
morin@irin.univ-nantes.fr

(3) Institut de Recherche en Communications et Cybernétique de Nantes - UMR CNRS La Chantrerie - Rue Christian Pauc - BP 50609 - 44306 Nantes Cedex 3
christian.viard-gaudin@polytech.univ-nantes.fr

Résumé – Abstract

Dans ce travail, nous étudions l'apport d'un modèle de langage pour améliorer les performances des systèmes de reconnaissance de l'écriture manuscrite en-ligne. Pour cela, nous avons exploré des modèles basés sur des approches statistiques construits par apprentissage sur des corpus écrits. Deux types de modèles ont été étudiés : les modèles n-grammes et ceux de type n-classes. En vue de l'intégration dans un système de faible capacité (engin nomade), un modèle n-classe combinant critères syntaxiques et contextuels a été défini, il a permis d'obtenir des résultats surpassant ceux donnés avec un modèle beaucoup plus lourd de type n-gramme. Les résultats présentés ici montrent qu'il est possible de prendre en compte les spécificités d'un langage en vue de reconnaître l'écriture manuscrite avec des modèles de taille tout à fait raisonnable.

This works highlights the interest of a language model in increasing the performances of on-line handwriting recognition systems. Models based on statistical approaches, trained on written corpora, have been investigated. Two kinds of models have been studied: n-gram models and n-class models. In order to integrate it into small capacity systems (mobile device), a n-class model has been designed by combining syntactic and contextual criteria. It outperforms bulkier models based on n-gram. The results we obtain show that it is possible to take advantage of language specificities to recognize handwritten sentences by using reasonable size models.

Mots Clés – Keywords

Reconnaissance de l'écriture manuscrite, modèle de langage, n-gramme, n-classe, perplexité.

Handwriting recognition, language modelling, n-gram, n-class, perplexity.

1 Introduction

Dans ce travail, nous nous intéressons au problème de la reconnaissance de l'écriture dite *en-ligne*. L'efficacité de celle-ci peut être renforcée à l'aide d'un modèle renfermant des connaissances *a priori* sur le langage. Dans un modèle probabiliste, une phrase s peut être représentée par une séquence de mots w_i de longueur L, soit : $s = w_1 \ldots w_i \ldots w_L = w_{1,L}$. En considérant cette séquence comme une chaîne de Markov, nous pouvons estimer la probabilité $p(s)$ d'une phrase s comme suit :

$$p(s) = p(w_{1,L}) = p(w_1)p(w_2|w_1)p(w_3|w_1 w_2)\ldots p(w_L|w_1 \ldots w_{L-1}) = \prod_{i=1}^{L} p(w_i|w_1 \ldots w_{i-1}) \quad (1)$$

Lorsque la longueur de l'historique du mot à prédire devient importante, l'estimation de la probabilité conditionnelle $p(w_i|w_1 \ldots w_{i-1})$ n'est pas fiable. La réduction de l'ordre de la chaîne de Markov permet alors de restreindre l'historique en ne tenant compte que du contexte proche des mots w_i (Manning et al., 2000). Dans un modèle n-gramme, seuls les $n-1$ précédents mots sont pris en considération :

$$p(w_i|w_{i-n+1} \ldots w_{i-1}) \approx p(w_i|w_1 \ldots w_{i-1}) \quad (2)$$

Les probabilités $p(w_i|w_{i-n+1} \ldots w_{i-1})$ sont calculées statistiquement par une simple méthode de comptage d'événements complétée par la méthode de lissage *absolute discounting backing-off* qui permet d'estimer $p(w_i|w_{i-n+1} \ldots w_{i-1})$ pour des évènements non rencontrés sur la base d'apprentissage. Pour évaluer l'adéquation du modèle, nous utilisons la traditionnelle mesure de perplexité :

$$PP_M(T_{test}) = \left[\prod_{i=1}^{L_{Test}} p(s)\right]^{-\frac{1}{L_{Test}}} \quad \text{(où } L_{Test} \text{ est le nombre de phrases dans la base de test)} \quad (3)$$

Comme le nombre de n-grammes devient vite considérable pour un lexique de taille importante dès lors que n augmente, nous cherchons à réduire le nombre d'événements observables en regroupant les mots en classes (nous parlerons alors de modèles n-classes). En appliquant ce regroupement, le modèle prédit non plus un mot en fonction des $n-1$ mots le précédant, mais en fonction des $n-1$ classes qui le précèdent.

$$p(s) = \sum_{che\min} \prod_{i=1}^{L} p(w_i|g_k(w_i))p(g_k(w_i)|g_{k'}(w_{i-1})g_{k''}(w_i)) \quad (4)$$

Dans le cas d'une classification « molle », chaque mot peut être associé à une ou plusieurs classes. Dans l'équation (4), un chemin correspond à une séquence $g(w_1)\ldots\ g(w_L)$ possible de classes. Dans le cas d'une classification « dure », chaque mot est associé à une et une seule classe; alors un seul chemin existe. Nous avons effectué la classification suivant deux critères. D'une part, un critère statistique regroupe les mots partageant les mêmes contextes lexicaux, et d'autre part, un critère syntaxique regroupe les mots selon leurs parties du discours.

2 Performances des modèles du langage

Dans cette section, nous cherchons à étudier le comportement de différents modèles de langage à savoir n-gramme, n-classe syntaxique et n-classe statistique sur différentes bases de données textuelles. Le tableau 1 présente les principales caractéristiques des corpus utilisés

pour l'apprentissage, puis pour l'évaluation des modèles. Pour des raisons de lisibilité, nous introduisons les abréviations suivantes : *MBG* pour Modèle BiGramme ; *MBCSynt* pour Modèle BiClasse Syntaxique et *MBCStatX* pour Modèle BiClasse Statistique utilisant X classes avec X={10, 50, 100, 500, 1000}.

	Nom	Taille du corpus (million de mots)	Taille du lexique (million de mots)	Domaine de discours	% mots avec une occurrence
Corpus d'apprentissage	ABU	4,1	0,09	romans du XIX^{e} et XX^{e}	42 %
	ECI	4,2	0,1	articles issus du journal Le Monde	42 %
Corpus de test	TEST	1,6	0,072	articles issus de journaux et de romans	44 %

Tableau 1 : Caractéristiques des corpus utilisés pour l'apprentissage et l'évaluation

2.1 Modèles n-classes statistiques et modèles n-classes syntaxiques

Dans le cas des modèles n-classes statistiques, nous utilisons un algorithme de classification dure inspiré de celui des *k*-means pour construire les classes, *k* étant le nombre de classes. Les résultats présentés à la figure 1 montrent, sans surprise, que plus le nombre de classes est important, meilleures sont les performances car les estimations de probabilités sont alors plus précises. Il est intéressant de noter que les MBCStat500/1000 parviennent presque à égaler les modèles bigrammes, ce qui est un résultat tout à fait intéressant étant donné leur moindre encombrement mémoire.

Dans le cas des modèles n-classes syntaxiques, nous utilisons des statistiques obtenues sur les corpus d'apprentissage préalablement étiquetés par l'étiqueteur de Brill (Brill 94 ; Le Comte et al., 1998) et de l'analyseur flexionnel Flemm (Namer 2000). Les performances de ces modèles qui comptent 210 classes sont inférieures à celles obtenues avec MBCStat500/1000 mais les MBCSynt prennent tout leurs intérêts lors de la combinaison avec d'autres modèles.

2.2 Combinaisons de modèles

A l'instar des travaux (Niesler, 1997 ; Jardino, 1994 ; Goodman, 2000 ; El-Bèze 1993), nous proposons d'étudier la combinaison linéaire de plusieurs de ces modèles, à savoir 1) modèles bigramme et biclasse syntaxique ; 2) modèles bigramme et biclasse statistique et enfin 3) modèles biclasse syntaxique et biclasse statistique. Par la suite, nous désignerons ces modèles sous le terme de modèle combiné. On peut observer sur la figure 2 les résultats correspondant à la combinaison des MBG et des MBCStat. Ces deux types de modèles apparaissent très complémentaires. Les MBCStat améliorent jusqu'à 18 % les performances du modèle bigramme sur ABU et jusqu'à 16 % sur ECI, pour les modèles avec 500 classes.

Le modèle biclasse syntaxique est également fortement complémentaire au modèle bigramme. Ainsi, la combinaison de ces deux modèles améliore de 45 % les performances sur ABU et de 33 % sur ECI. Les MBCStat500 présentaient de meilleures performances que les MBCSynt, or la combinaison entre un MBG et un MBCStat500 est moins performante qu'une combinaison entre un MBG et un MBCSynt. On peut en conclure que la nature des informations contenues dans les MBCSynt est plus complémentaire de celle des MBG que ne

l'est celle des MBCStat. Un MBCStat est par nature proche des MBG et apporte donc moins qu'un MBCSynt à un MBG.

Enfin, dans le cas d'une combinaison des modèles biclasse statistique et biclasse syntaxique, laquelle configuration correspond à une combinaison réellement intéressante car seuls des modèles de faibles complexités sont pris en compte, les performances sont notablement améliorées (cf. figure 2). Ses résultats surpassent très significativement ceux obtenus par des MBG. On obtient jusqu'à 47 % d'amélioration sur ABU et 35 % sur ECI avec un MBCSynt combiné à un MBCStat500 par rapport au MBG. Toutefois, on peut conjecturer que la relative faiblesse des modèles bigrammes par rapport aux modèles n-classes combinés est en partie due à la taille réduite de la base d'apprentissage (4 millions contre plusieurs dizaines de millions de mots pour obtenir des modèles bigrammes véritablement robustes).

Nous avons enfin essayé de combiner les trois modèles : MBG, MBCStat1000 et MBCSynt. Les performances obtenues ne sont que très légèrement supérieures (1 à 2 % sur ABU et 5 % sur ECI) à celles correspondant à la combinaison des modèles biclasses seuls. Les MBG n'apportent donc que peu d'informations complémentaires aux modèles MBCStat et MBCSynt combinés. On peut en conclure que les modèles combinés MBCStat et MBCSynt se suffisent à eux-mêmes et ne nécessitent pas de combinaisons avec un MBG ce qui de toutes manières constituerait alors un modèle beaucoup trop volumineux.

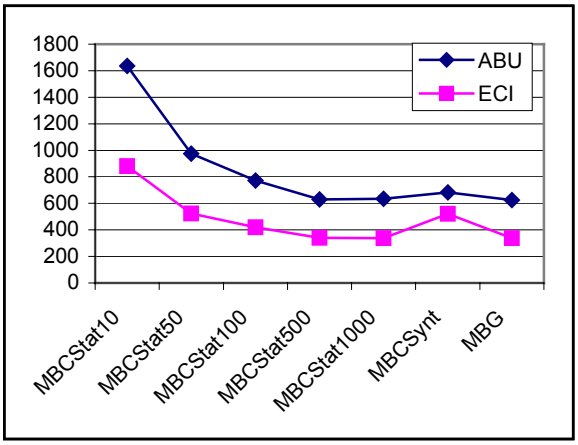

Figure 1 : Mesures de perplexité avec les modèles simples sur le corpus TEST

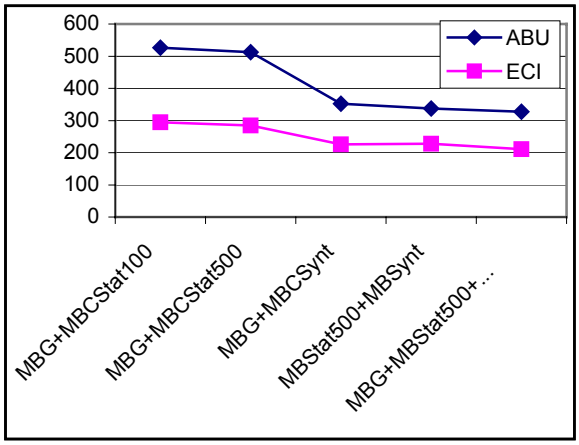

Figure 2 : Mesures de perplexité avec les modèles combinés sur le corpus TEST.

3 Contribution des modèles de langage dans le système de reconnaissance

Afin d'évaluer la contribution des modèles du langage dans le système de reconnaissance de l'écriture, nous avons utilisé une nouvelle base de test composée de 4 912 phrases distinctes[1].

[1] Cette base de test, composée de 37 700 mots définissant un lexique de 8 800 mots, est issue d'une collecte réalisée à l'aide d'une ardoise électronique auprès de 400 scripteurs.

La figure 3 présente un exemple du signal d'entrée du système correspondant aux phrases « *Mais jamais pour très longtemps.* » et « *Ce n'est pas si sûr.* ».

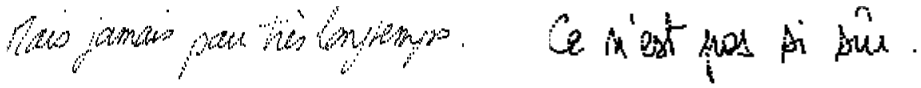

Figure 3 : Echantillons de la base d'écriture manuscrite dynamique

Sur cette base, le taux d'erreur de reconnaissance mot, sans aucun modèle de langage est égal à 34 %. Cette valeur est bien supérieure à ce que l'on obtient plus classiquement sur une base de mots. Il faut souligner en particulier ici le fait que, outre la difficulté de la segmentation inter-mot qui n'existe pas dans une reconnaissance base mot, la ponctuation est prise en compte dans la reconnaissance (par exemple chaque erreur sur une virgule est comptabilisée). Si l'on rajoute un modèle de langage élémentaire, consistant en l'intégration des monogrammes dans le treillis de reconnaissance, alors le taux d'erreur chute à 29 %.

La figure 4 présente l'évolution du taux d'erreur avec des systèmes de reconnaissance intégrant des modèles monogramme et des modèles biclasse, biclasse combiné ou bigramme. La diminution du taux d'erreur est très significative. Pour le MBCStat500 seul, le taux d'erreur est inférieur à 23 %, cela correspond à une diminution de l'erreur de 32 % par rapport au système sans modèle de langage. Le meilleur résultat est obtenu avec la combinaison des modèles MBCStat500 et MBCSynt où le taux d'erreur est de 22,5 %.

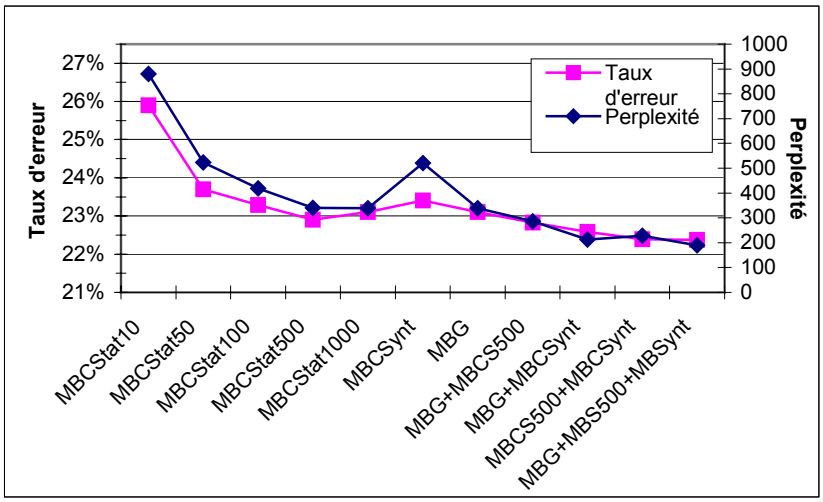

Figure 4 : Mesures de perplexité et taux d'erreur avec le modèle entraîné sur le corpus ECI.

De plus, ces courbes mettent clairement en évidence la forte corrélation entre la mesure de perplexité et le taux d'erreurs. La mesure de perplexité semble donc bien un indicateur valide pour mesurer la pertinence d'un modèle de langage en vue de son utilisation dans un contexte de reconnaissance de l'écriture manuscrite. Elle a l'avantage de pouvoir être évaluée sur des bases beaucoup plus faciles à obtenir que des bases d'écriture manuscrite. Ce point est important et n'avait pas à notre connaissance été préalablement mis en évidence expérimentalement comme nous le faisons ici.

Si nous reprenons les exemples de la figure 3, le système de reconnaissance sans modèle de langage propose en sortie : « *Mais jamais pour **tirs** longtemps.* » et « *Ce n'est pas si **dû***. ». Par exemple, le MBCSynt corrige l'erreur sur ces deux exemples. D'un point de vue syntaxique,

il est effectivement plus vraisemblable que l'adverbe de temps « *longtemps* » soit précédé d'un adverbe que d'un nom commun. De même, après l'adverbe « *si* », l'adjectif « *sûr* » est plus probable que le participe passé « *dû* ».

4 Conclusion et perspectives

Dans ce travail, nous avons montré l'apport significatif des modèles de langage à un système de reconnaissance de l'écriture manuscrite en ligne. Globalement, nous avons obtenu une diminution de plus 34 % du taux d'erreur en recherchant le meilleur compromis performance/coût matériel.

Il reste bien entendu un certain nombre de points à améliorer. Tout d'abord, il serait intéressant d'étudier les techniques de classification permettant de travailler sur des corpus plus volumineux (Beaujard et *al.*, 1999 ; Goodman, 2000). Ensuite, nous devons affiner nos modèles qui souffrent actuellement d'un manque de robustesse lorsqu'ils sont confrontés à des noms propres ou à des mots d'origine étrangères. Enfin, nos applications étant destinées à des appareils nomades que l'utilisateur s'approprie, il serait intéressant d'adapter nos modèles au domaine de discours de l'utilisateur.

Références

Beaujard C., Jardino M., *Classification de mots non étiquetés par des méthodes statistiques*, *Mathématiques informatique et Sciences Humaines*, vol 147, pp. 7-23, 1999.

El-Bèze M., *Les modèles de langage probabilistes : quelques domaines d'application*, HDR, LIPN, 1993.

Brill E., *Some Advances in Rule-Based Part of Speech Tagging*, In Proceedings, Twelfth National Conference on Artificial Intelligence (AAAI'94), pp 722-727, 1994.

Goodman J., *Putting it all together : Language model combination*, ICASSP-2000, Istanbul, 2000.

Jardino M., *Automatic determination of a stochastic bigram class model*, International Colloquium on Grammatical Inference, 1994.

Le Comte J., Paroubek P., *Le catégoriseur d'Eric BRILL. Mise en oeuvre de la version entraînée à l'INALF*, Rapport technique, Nancy, CNRS-InaLF, 1998.

Manning C., Scutze., H. *Foundation of Statistical Natural Language Processing*, The MIT Press, 2000.

Namer F., *Flemm : Un analyseur flexionnel du français à base de règles, Traitement Automatique des Langues (TAL)*, 41(2) pp. 523-548, 2000.

Niesler T., *Category Based Statistical Language Models*, Ph. D. thesis, University of Cambridge, June 1997.