Utilisation de la classification K-means pour les modèles de langage à base de classes morphologiques

A. Ghaoui[(1) (2) (3)], F. Yvon[(2)], C. Mokbel[(1)] et G. Chollet[(2)]

[(1)] Universite de Balamand
BP 100 Tripoli Liban
Chafic.Mokbel@balamand.edu.lb
(2) CNRS-URA820, Ecole Nationale Supérieure des Télécommunications
46 rue Barrault, 75634 Paris cedex 13, France
Francois.Yvon@enst.fr, Gerard.Chollet@enst.fr
(3) Jinny Software Ltd, Samra Center, 1st floor, Fanar, El Metn, Lebanon
Antoine.Ghaoui@jinny.ie

ABSTRACT

The introduction of morphological constraints in the language modelling for Arabic language is of particular of interest. Previously proposed morphological class-based N-gram models have shown satisfactory results. This model had a limitation because morphological rules have been only related to the contextual roots of the previous words. In the present paper this limitation has been removed without increasing the number of parameters. This has been made possible by classifying the contexts of the morphological rules using the k-means algorithm. The language model has been experimented and the measured perplexity shows satisfactory results especially when the classical N-gram is interpolated with the new model. This interpolation relatively improves 13% the perplexity with respect to the N-gram.

1. INTRODUCTION

Les modèles de langage statistiques sont largement utilisés notamment dans des systèmes de reconnaissance de la parole. Ces modèles sont basés sur la détermination de la distribution de probabilité des mots conditionnellement au contexte gauche. Aucune distribution paramétrique n'existe jusque là dû à l'absence de relations d'ordre dans l'espace des mots du vocabulaire. Ainsi, le nombre de paramètres d'une telle modélisation croît exponentiellement avec la taille du vocabulaire. Ceci implique une complexité au niveau de l'apprentissage surtout pour les contextes rarement rencontrés en pratique. Ce problème apparaît d'autant plus lorsque l'on cherche à adapter un modèle de langage à un type de langage particulier. Pour cela, des recherches ont été effectuées pour factoriser les modèles de langage en introduisant une classification P. F. Brown, V.J. Della Pietra, P.V. deSouza, J.C. Lai, & R.L. Mercer [3]. Les modèles à base de classes ont été appliqués avec succès en adaptation.

Dans A. Ghaoui, F. Yvon, C. Mokbel. & G. Chollet [9][10] nous avons proposé d'introduire la classification morphologique afin de factoriser la modélisation du langage. Pour des langues à riches structures morphologiques, l'idée consiste à séparer la dépendance entre les mots en deux dépendances, une entre les racines des mots et l'autre entre les règles morphologiques et les racines. Cette simplification nous a permis de réduire largement le nombre de paramètres au prix d'une perte de la perplexité. L'analyse expérimentale nous a conduit à identifier la faiblesse de l'approximation, à savoir négliger la dépendance entre les règles morphologiques (par exemple, négliger le lien des règles morphologiques du pluriel dans 2 mots consécutifs: "الولدان المجتهدان" ou *jeune***s** *enfant*s). Pour remédier au problème il suffit de reintroduire la dépendance entre les règles morphologiques et les règles du contexte. Ceci augmenterait significativement le nombre de paramètres. Nous proposons ainsi de factoriser les contextes de distributions conditionnelles des règles en utilisant l'algorithme des K-means. Les expériences conduites montrent une amélioration significative de la perplexité en réduisant le nombre de paramètres. De plus, la combinaison avec le modèle N-gram d'origine permet une nette amélioration de la perplexité.

La section suivante décrit le modèle base de contraintes morphologiques et l'introduction de la classification par K-means des contextes des règles morphologiques. Les experiences sur une base de données en langue arabe et les résultats obtenus sont détaillées dans la section 3. Des conclusions et perspectives terminent le papier.

2. MODÈLE DE LANAGUAGE A BASE DE CONTRAINTES MORPHOLOGIQUE

Dans une langue, beaucoup de mots sont dérivés à partir de racines. L'introduction de contraintes morphologiques dans la modélisation du langage prend un intérêt particulier. Ceci est d'autant plus vrai pour des langues telles que l'Arabe et l'Allemand.

Soient w_i le i^{ème} mot du vocabulaire, r_i la racine de ce mot et g_i la règle morphologique qui permet la dérivation du mot w_i à partir de r_i. Ainsi, chaque racine possible dans le vocabulaire définit une classe dans laquelle se retrouvent tous les mots du vocabulaire se dérivant de cette racine.

Considérant un modèle N-gram où pour un contexte de N-1 mots on définit la distribution discrète des mots du vocabulaire V:

$\{\Pr(w_n/w_{n-1}, w_{n-2}, \ldots, w_{n-N+1})\}$ où $w_i \in V$

Chaque mot w_i s'écrit comme le couple (r_i, g_i). Ainsi la probabilité d'un mot sachant le contexte s'écrit:

$$\Pr(w_n/w_{n-1}, \ldots, w_{n-N+1}) = \Pr((r_n, g_n)/(r_{n-1}, g_{n-1}), \ldots, (r_{n-N+1}, g_{n-N+1}))$$

$$= \Pr(r_n, g_n/r_{n-1}, g_{n-1, \ldots,} r_{n-N+1}, g_{n-N+1})$$

$$= \Pr(g_n/r_n, r_{n-1}, g_{n-1}, \ldots, r_{n-N+1}, g_{n-N+1}) \Pr(r_n/r_{n-1}, g_{n-1}, \ldots, r_{n-N+1}, g_{n-N+1}) \quad \text{(EQ.1)}$$

Le premier lissage des distributions qui peut se faire est que la probabilité de la $n^{\text{ième}}$ racine r_n ne dépend que des racines précédentes $(r_{n-1}, \ldots, r_{n-N+1})$. L'Eq. 1 devient:

$$\Pr(w_n/w_{n-1}, \ldots, w_{n-N+1}) \cong \Pr(g_n/r_n, r_{n-1}, g_{n-1}, \ldots, r_{n-N+1}, g_{n-N+1}) \Pr(r_n/r_{n-1}, \ldots, r_{n-N+1}) \quad \text{(EQ. 2)}$$

L'Eq. 2 définit un premier modèle à base de classes morphologiques. Supposant que le nombre de racines pour le vocabulaire V de taille v est n_r, et que le nombre de règles possible est n_g (de l'ordre de quelques centaines); le nombre total de paramètres dans le modèle de l'Eq. 2 est $n_g n_r (n_r n_g)^{N-1} + n_r^N$. Ce modèle n'offre pas de réduction du nombre de paramètres par rapport au modèle N-grams. Dans un travail précédent A. Ghaoui, F. Yvon, C. Mokbel. and G. Chollet [9][10] nous avons introduit d'autres approximations en limitant la dépendance d'une règle g_n aux racines $\Pr(g_n/r_n, r_{n-1}, g_{n-1}, \ldots, r_{n-N+1}, g_{n-N+1}) \cong \Pr(g_n/r_n, r_{n-1}, \ldots, r_{n-N+1})$. Cette approximation a déterioré la modélisation.

2.1. FACTORISATION DU MODELE DE REGLES PAR L'ALGORITHME K-MEANS

Pour garder l'Eq. 2 tout en réduisant le nombre de paramètres nous proposons dans ce travail de factoriser les distributions du type $\Pr(g_n / r_n, r_{n-1}, g_{n-1})$. Cette factorisation est faite en utilisant l'algorithme des K-means. Ceci revient a trouver une fonction de classification $C(r_n, r_{n-1}, g_{n-1})$ qui permet une bonne approximation:

$$\Pr(g_n/r_n, r_{n-1}, g_{n-1}) \cong \Pr(g_n/C(r_n, r_{n-1}, g_{n-1})) \quad \text{(EQ 3)}$$

Il s'agit de déterminer K classes de contextes de manière à conserver la meilleure description par le modèle de langage à base de règles. Le nouveau modèle devient :

$$\Pr(w_n/w_{n-1}, \ldots, w_{n-N+1}) \cong \Pr(g_n/C(r_n, r_{n-1}, g_{n-1}, \ldots, r_{n-N+1}, g_{n-N+1}) \Pr(r_n/r_{n-1}, \ldots, r_{n-N+1}) \quad \text{(EQ 4)}$$

Dans ce cas, le nombre de paramètres est réduit à : $n_g k + n_r^N$.

n_g est de l'ordre de quelque centaines, k est de quelques dizaines, ce qui donne un nombre de paramètre $\cong n_r^N$. Ce nouveau modèle offre donc un lissage de $O(n_g^N)$.

L'algorithme prend en entrée le modèle général $\Pr(g_n / r_n, r_{n-1}, g_{n-1})$ et le nombre de classes K. Il procède comme suit:

1. Initialisation: répartir les triplets contextes (r_n, r_{n-1}, g_{n-1}) sur K classes d'une manière aléatoire. Chaque triplet est ensuite étiqueté par sa classe.
2. Itérations:
   a. Pour chaque classe estimer la distribution $\Pr[g_n / C(r_n, r_{n-1}, g_{n-1})]$ et calculer la perplexité sur des données d'apprentissage ou de développement.
   b. Pour chaque contexte trouver la meilleure classe selon le critère de divergence de Kullback-Leibler.

   Soient
   $f_1(g_n) = \Pr[g_n / r_n, r_{n-1}, g_{n-1}]$
   et
   $f_2(b_n) = \Pr[g_n / C(r_n, r_{n-1}, g_{n-1})]$
   les distributions pour un contexte et pour une classe de contextes, et B le nombre de règles. On attribue la classe la plus proche en utilisant la divergence de KL:
   $$d(f_1, f_2) = \sum_{k=1}^{B} f_1(g_k) \log \frac{f_1(g_k)}{f_2(g_k)}$$
   c. Recommencer (2) jusqu'à la stabilité (c'est à dire aucun contexte ne change de classe dans (2.b) ou jusqu'à le nombre maximum d'itérations est atteint.
3. Remplacer pour chaque contexte sa distribution par la distribution de la classe correspondante.

2.2. TRANSDUCTEUR MORPHOLOGIQUE

L'analyse morphologique de l'Arabe est un domaine bien étudié. Des techniques à base de transducteurs ont été développées K. Beesley [6]. Les verbes et les mots en Arabe sont généralement déduits d'un nombre limité de racines, il est intéressant d'essayer de retrouver les racines de chacun des mots du vocabulaire. Dans un travail précédent A. Ghaoui, F. Yvon, C. Mokbel. and G. Chollet [9][10], nous avons proposé des automates décrivant des règles simples et élémentaires. Les mots sont des racines auxquelles sont associés des préfixes et des suffixes. Les préfixes et les suffixes sont fixés.

3. EXPÉRIENCES ET RÉSULTATS

Les expériences sont menées sur la base de données des articles du journal Al-Nahar pour les deux années consécutifs 1998 et 1999 et en utilisant SRILM. Dans ce suit nous présentons une brève description de la base de données, les expériences menées et les résultats obtenus en terme de perplexité.

3.1. Base de données

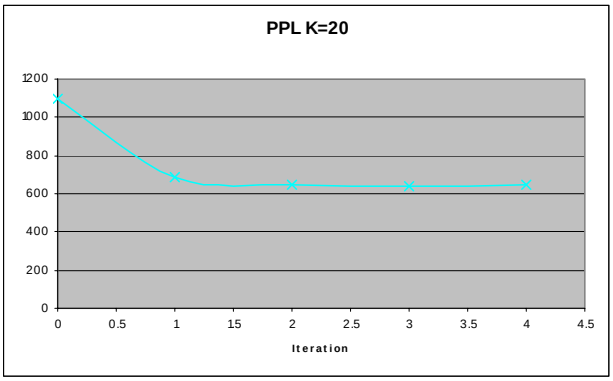
Les expériences ont été menées sur les articles du journal Al-Nahar pour les années 1998 et 1999. La base de données est divisée en deux parties, l'année 1999 pour l'apprentissage et l'année 1998 pour le test. La partie apprentissage qui correspond à l'année 1999 est formée de 429229 mots différents avec une fréquence d'apparition qui varie entre 1 et 754100. Quand à la partie test, elle est constituée de 440298 mots différents avec une fréquence d'apparition qui varie entre 1 et 733023.

Dans nos expériences préliminaires, le vocabulaire a été limité aux mots dont la fréquence d'apparition est supérieure à 100. Ceci produit un vocabulaire de 18119 mots différents. L'analyse morphologique donne 10269 racines différentes et 198 règles différentes.

3.2. Implémentation

L'implémentation du modèle proposé à demandé le développement d'un module qui permet de générer a partir du modèle $\Pr(g_n / r_n, r_{n-1}, g_{n-1})$ et le nombre de class le modèle $\Pr(g_n / C(r_n, r_{n-1}, g_{n-1}))$. Le N-gram sur les racines des mots est fait en utilisant le logiciel SRILM et l'analyseur morphologique. Une fois le modèle N-gram sur les racines et le modèle des règles a base de classe sont calculés, le modèle est appliqué sur les données de test. Le code du SRILM A. Stockle [1] a été modifié pour considérer la spécificité du modèle l'EQ. 4. Pour chaque mot et chaque contexte, les racines, les règles et les classes correspondantes sont retrouvées. La probabilité N-gram sur ces racines est ensuite déterminée comme usuellement dans SRILM. Ensuite cette probabilité est multipliée par la probabilité de la règle qui a permis de retrouver la racine pour le mot courant sachant la classe de la racine courante, la racine précédente et la règle du mot précédent.

Les expériences sont faites sur plusieurs valeurs de k (2, 5, 10, 20). Pour toutes les valeurs de K, l'algorithme converge très rapidement et la valeur de perplexité atteint une valeur optimale. Le graphe suivant montre la convergence de la valeur de la perplexité pour k=20.

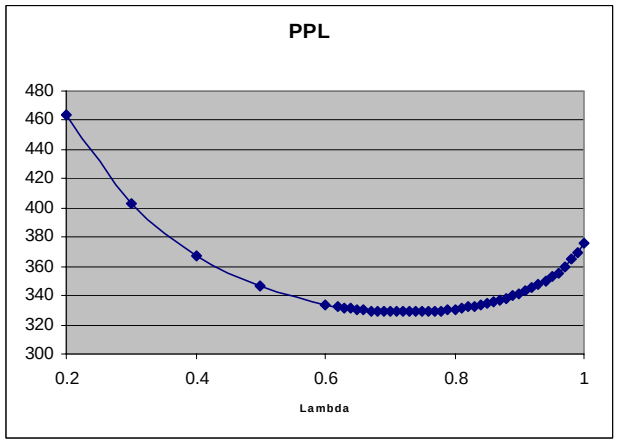


3.3 Résultats

La perplexité calculée pour le N-gram classique est égale à 376.213. La modèle morphologique à classification par K-means donne pour une valeur de k = 20 une perplexité de 640.576. Cette valeur de perplexité comparée avec les valeurs de perplexité obtenues sur les mêmes données en utilisant différentes modèles A. Ghaoui, F. Yvon, C. Mokbel. and G. Chollet [9][10] montre que le modèle a base de classes morphologiques et utilisant les k-means donne des meilleurs résultats.

3.4. Interpolation linéaire avec le modèle N-gram classique

L'interpolation linéaire entre le modèle proposé après classification en utilisant k-means et le modèle n-gram de base donne une amélioration en termes de perplexité de 13% avec une valeur minimale de 328.847. Ceci montre que notre modèle apporte une information complémentaire au modèle N-gram classique.



4. CONCLUSIONS ET PERSPECTIVES

Dans un travail précédent nous avons proposé un modèle de langage statistique de type N-gram introduisant des classes morphologiques. Les approximations effectuées pour réduire le nombre de paramètres consistaient à éliminer la dépendance entre règles morphologiques ce qui a réduit la précision de la modélisation. Enlever cette approximation augmenterait significativement le nombre de paramètres et fait perdre l'utilité de l'approche par classes morphologiques. Dans ce papier le problème est résolu en factorisant le contexte de dépendance d'une règle morphologique. Cette factorisation est effectuée par l'algorithme K-means. Les premières expériences ont montré une nette amélioration de la perplexité. De plus la combinaison linéaire de ce modèle avec le modèle N-gram classique permet une amélioration de l'ordre de 13% de la perplexité. Ceci démontre que le modèle proposé apporte une complémentarité par rapport au modèle N-gram classique.

Dans ce qui suit, l'étude sera poursuite pour optimiser le nombre de classes recherchées par K-means. De plus, une étude du résultat de la classification sera effectuée pour

comprendre ses effets en fonction des mots. Il s'agit de comprendre si des contextes peuvent produire des règles morphologiques similaires.

BIBLIOGRAPHIE

[1] A. Stockle SRILM - An Extensible Language Modeling Toolkit *Proc. Intl. Conf. Spoken Language Processing*, Denver, Colorado 2002.

[2] C. Manning and H. Schutze *Foundations of Statistical Natural Language Processing*, MIT Press. Cambridge 1999.

[3] P. F. Brown, V.J. Della Pietra, P.V. deSouza, J.C. Lai, and R.L. Mercer Class-based N-gram models of natural language, *Computational Linguistics*, Vol. 18, pp. 467–479 1992

[4] T. Neisler *Category-based statistical language models* Ph.D thesis, Department of Engineering, University of Cambridge, U.K., 1997

[5] K. Darwish, Building a Shallow Arabic Morphological Analyzer in One Day *Proceedings of the ACL Workshop on Computational Approaches to Semitic Languages*, Philadelphia, Boston, 2002.

[6] K. Beesley Arabic Finite-State Morphological Analysis and Generation *COLING-96*, Vol. 1, pp. 89-94 1996

[7] S. Katz Estimation of Probabilities from Sparse Data for the Language Model Component of a speech Recognizer *IEEE Trans. on Acoustics, Speech, and Signal Processing*, Vol. ASSP-35, no. 3, pp. 400-401 MARCH 1987

[8] M. Jardino and G. Adda. Language modelling for CSR of large corpus using automatic classification of words *Proc. European Conference on Speech Communication and Technology*, 1191--1194, September 1993.

[9] A. Ghaoui, F. Yvon, C. Mokbel and G. Chollet Modèle de langage statistique à base de classes morphologiques, *Le traitement automatique de l'arabe JEP-TALN 2004*, Fès, 19-22 avril 2004.

[10] A. Ghaoui, F. Yvon, C. Mokbel. and G. Chollet On the Use of Morphological Constraints in N-gram Statistical Language Model *Interspeech, 9th European Conference on Speech Communication and technology*, Sept. 4-8 Lisbon, Portugal 2005