Modèle de langage statistique à base de classes morphologiques

A. Ghaoui(1)(2), F. Yvon(2), C. Mokbel(1) et G. Chollet(2)
(1) Université de Balamand
BP 100 Tripoli, Liban
Antoine_Ghaoui@hotmail.com Chafic.Mokbel@balamand.edu.lb
(2) CNRS-URA820, Ecole Nationale Supérieure des Télécommunications
46 rue Barrault, 75634 Paris cedex 13, France
{francois.yvon, gerard.chollet}@enst.fr

Résumé – Abstract

Dans la langue Arabe, beaucoup de mots sont dérivés à partir de racines. Ainsi, l'introduction de contraintes morphologiques dans la modélisation du langage prend un intérêt particulier. Cet article décrit un modèle de langage statistique à base de classes morphologiques, dont chaque racine forme une classe. L'efficacité des modèles à base de classes a été prouvée. Ces modèles apportent un plus pour l'adaptation et pour l'apprentissage de modèles à partir de bases de données réduites, et ils nécessitent généralement un espace mémoire réduit. Dans le cadre de modèles à base de classes, ce papier propose un cadre pour l'utilisation de classes morphologiques. De plus, un modèle morphologique pour la langue Arabe à base de règles d'extraction des racines est aussi proposé. Des expériences préliminaires ont été conduites sur une base de données formée des articles du Journal Al-Nahar pour les années 1998 et 1999. Les résultats sont donnés dans ce papier.

In Arabic language a lot of words are derived from their root. Thus, the introduction of morphological constraints in the language modelling is of particular interest. Class-based N-gram models have shown satisfactory results especially for language model adaptation and for training of reduced datasets. They have also been very effective in their use of the memory space. This paper describes a language model using word classes where each root is a class. In addition, a rule-based stemming method is also proposed for the Arabic language. The language model has been experimented on a database formed of the Al-Nahar newspaper articles for the years 1998 and 1999. Preliminary results are given in this paper.

Keywords – Mots Clés

Modèles statistiques de langage, N-grams, modèles à base de classes, contraintes morphologiques.
Statistical language models, N-grams, class-based models, morphological constraints.

1 Introduction

Les modèles de langage les plus couramment utilisés sont purement statistiques. Les systèmes de reconnaissance de parole continue à l'état de l'art intègrent les N-grams [Manning 1999] comme modèles statistiques du langage. Un modèle N-grams définit le langage comme une distribution discrète de probabilités d'observer un mot sachant qu'on connaît les N-1 mots précédents. Ces modèles statistiques sont à la base caractérisés par un nombre de paramètres très large. En effet, pour un vocabulaire de V mots, un simple bigram aurait V^2 paramètres où le paramètre (i,j) représente la probabilité que le mot i suit le mot j. Ceci pose un double problème. Premièrement, l'apprentissage nécessite énormément de données et il est difficile de s'assurer, même sur de grosses bases de données, que toutes les combinaisons de mots apparaissent. Deuxièmement, l'utilisation de ces modèles nécessitent énormément d'espace mémoire.
Afin de résoudre ces problèmes, il est nécessaire de paramétrer les distributions discrètes sur l'espace des mots du vocabulaire. Cependant, une telle approche nécessiterait une norme ou du moins une notion d'ordre sur l'ensemble des mots du vocabulaire. Or ces relations n'existent pas. D'où l'introduction d'approches à base de classes qui exploitent la classification des mots du vocabulaire pour lisser et paramétrer les distributions discrètes des modèles N-grams [Brown 1992]. Les classes utilisées en général sont de nature syntaxique, sémantique ou morphologique ou des classes déterminées automatiquement. Dans ce travail, nous nous intéressons en particulier aux classes morphologiques.
Pour définir des classes morphologiques, on considère qu'un mot du vocabulaire peut être généré à partir de sa racine par l'application d'une règle morphologique. La probabilité d'un mot sachant son contexte peut être ainsi déduite d'un modèle de langage type N-grams sur les racines des mots avec un ajustement selon la règle de production à partir de ces racines. Le cadre théorique formalisant cette idée est donné dans la section suivante.
La section 3, définit un modèle morphologique où un transducteur est utilisé pour décrire un ensemble de règles. Ces règles sont définies d'une manière empirique. Ce transducteur est intégré dans notre modélisation du langage.
Une version très simplifiée de notre modèle de langage à base de contraintes morphologiques a été implantée dans le logiciel SRILM [Stolcke 2002] et dans le logiciel CMU Toolkit [Clarkson 1997]. Des expériences préliminaires sont conduites sur une base de données formée des articles du journal Al-Nahar pour les années 1998 et 1999. La base de données, les expériences et les résultats obtenus sont décrits dans la section 4. Finalement, le papier se termine par des conclusions et des perspectives.

2 Modèle de langage à base de classe morphologique

La modélisation de langage par *N*-gram intégrant des classes de mots du vocabulaire est bien étudiée [Manning 1999]. Dans ce qui suit une approche particulière est proposée où des règles morphologiques sont utilisées pour la définition des classes.
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:

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

La première approximation ou 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 classique.

La seconde approximation que l'on peut faire est que la règle morphologique à l'instant *n* ne dépend pas des racines aux instants précédents. On déduit de l'Eq. 2:

$$\Pr(w_n/w_{n-1}, \ldots, w_{n-N+1}) \cong \Pr(g_n/ r_n, g_{n-1}, \ldots, g_{n-N+1}) \Pr(r_n/r_{n-1}, \ldots, r_{n-N+1}) \quad \text{(EQ. 3)}$$

Ce nouveau modèle est caractérisé par un nombre de paramètres égal à:

$n_g n_r (n_g)^{N-1} + n_r^N = n_r n_g^N + n_r^N$.

En comparant au nombre de paramètres v^N du modèle N-gram et en admettant les hypothèses suivantes :

- $v = n_g n_r$,
- $nr >> ng$

On peut conclure que le nouveau modèle de l'Eq. 3 offre un lissage du modèle N-gram d'origine par un facteur $O(n_g^N)$.

On peut aussi pousser l'approximation en acceptant que la règle ne dépend que du type de la racine, à savoir: verbe, nom, adverbe, prénom, etc. Ce qui nous conduit à:

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

Si finalement on néglige la dépendance de la règle de l'instant *n* aux règles des instants précédents, nous obtenons le modèle:

$$\Pr(w_n/w_{n-1}, \ldots, w_{n-N+1}) \cong \Pr(g_n/ T(r_n)) \Pr(r_n/r_{n-1}, \ldots, r_{n-N+1}) \quad \text{(EQ. 5)}$$

Ce modèle est celui que nous avons commencé à expérimenter dans ce travail. Nous avons considéré jusqu'à présent deux types de racines : verbe et non-verbe. $\Pr(g_n/ T(r_n))$ est alors calculé comme deux distributions des règles morphologiques qui distinguent entre le type des racines, les racines verbes et les racines non verbes.

3 Transducteur Morphologique

Sachant que 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.

L'analyse morphologique de l'Arabe est un domaine bien étudié. Des techniques à base de transducteurs ont été développées [Beesley 1996]. Dans ce travail nous proposons 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. Ainsi les règles se résument à:

- Mots qui commencent par AL (‘ال’, ‘وال’, ‘فال’, ‘بال’, ‘كال’, ‘لل’).
- Mots qui commencent par AL et qui se terminent par PLUR (‘ان’, ‘ين’, ‘ون’).
- Mots qui commencent par GEN (‘و’, ‘ف’, ‘ب’, ‘ك’, ‘ل’, ‘س’) et qui se terminent par POS (‘تموني’, ‘تمونا’, ‘تهما’, ‘وهما’, ‘تينا’, ‘تيني’, ‘تمونِي’, ‘تكَ’, ‘تكِ’, ‘تني’, ‘وهم’, ‘وهن’ , ‘تنا’, ‘تكم’, ‘تكن’, ‘تهم’, ‘تهن’, ‘تها’, ‘ وكَ’, ‘ وكِ’, ‘وكم’, ‘وكن’, ‘وها’, ‘ونِي’, ‘ونا’, ‘هما’, ‘ تً’, ‘ تِ’, ‘ك’, ‘ك’, ‘كم’, ‘كن’, ‘ني’, ‘ها’, ‘هم’, ‘هن’, ‘وا’, ‘وك’, ‘وه’, ‘تك’, ‘ته’, ‘تم’, ‘نا’, ‘ تَ’, ‘ك’, ‘ه’, ‘ي’, ‘ت’).
- Mots qui commencent par GEN et qui se terminent par PLUR.
- Mots qui se terminent par POS.
- Mots qui se terminent par PLUR.

Afin d'illustrer ce modèle considérons l'exemple de la règle AL+*racine* suivant:

“الولد” (l’enfant) = "ولد+ال" (l’+enfant)

L'implémentation de l'extraction des racines des mots est faite en utilisant une structure de graphe transducteur (‘‘Finite State Tranducer’’ (FST)) comme le montre la figure 1. Huit classes de règles existent au total :

- racine
- AL_racine
- AL_racine_PLUR
- GEN_racine
- GEN_racine_POS
- GEN_racine_PLUR
- racine_POS
- racine_PLUR

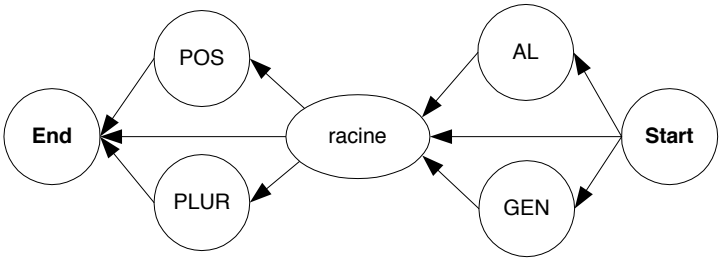

Figure 1 : Transducteur implémentant les règles morphologiques.

Pour le calcul de la distribution discrète $\Pr(g_n/\mathrm{T}(r_n))$, on considère qu'elle est indépendante du mot et de sa racine, et ne dépend que de la règle de passage utilisée et du type de la racine. Deux cas sont considérés pour le type de la racine. Premièrement, toutes les racines sont du même type et une seule et unique distribution générale des règles est calculée. Deuxièmement, les racines sont divisées en verbes et non-verbes. Un dictionnaire de 10000 verbes est utilisé pour effectuer cette distinction. Deux distributions discrètes sur les règles sont ensuite déterminées.

4 Expériences et Résultats

Les expériences sont effectuées en utilisant le logiciel SRILM [Stolcke 2002] et la base de données des articles du journal Al-Nahar pour deux années consecutives, à savoir 1998 et 1999. Le logiciel SRILM n'intègre pas une solution qui permet d'expérimenter le modèle proposé. Ainsi, ce modèle a été intégré dans le logiciel. Dans ce qui suit nous donnons une

brève description de la base de données et de l'implémentation dans SRILM du modèle proposé. Les résultats expérimentaux sont ensuite détaillés.

4.1 Base de données Al-Nahar

Les expériences ont été menées sur les articles du journal Al-Nahar pour les années 1998 et 1999. Dans ce travail, la base de données a été divisée en deux parties, l'année 1999 pour l'apprentissage et l'année 1998 pour le test. Les articles se présentent sous la forme de pages HTML. Un outil de prétraitement a été développé pour extraire le texte de ces pages. La partie apprentissage qui correspond à l'année 1999 est formée de 44234 pages HTML. La partie test est formée de 47766 pages. Après le prétraitement, la partie apprentissage de la base de données contient 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 qui correspondent à 10269 racines différentes.

4.2 Implémentation dans SRILM

SRILM [Stolcke 2002] implémente une technique de modélisation du langage à base de classes. Son utilisation pour introduire des classes morphologiques s'est avérée complexe vu le nombre de classes qui est égal au nombre de racines. Ainsi le logiciel a été modifié pour introduire le modèle proposé dans ce papier.
L'idée est d'utiliser le logiciel SRILM pour calculer un N-gram sur les racines des mots. Une fois ce N-gram défini, les paramètres du backoff sont recalculés. Le backoff est une technique qui consiste à lisser le modèle du langage. Si l'on ne trouve pas d'occurrences pour un mot dans un contexte donné, on regarde dans le contexte plus large précédent jusqu'à ce que nous trouvons une occurrence ou que l'on arrive à l'unigram. La probabilité trouvée au contexte plus large, est normalisée par un facteur linéaire. La technique du backoff classique ne s'applique pas dans notre cas car les probabilités sont calculées selon l'Eq. 5. Ceci a nécessité la modification du code afin de tenir compte du modèle de l'Eq. 5 dans le calcul.
Une fois le modèle N-gram sur les racines calculé, le modèle est appliqué sur les données de test. Là aussi le code SRILM a été modifié pour considérer la spécificité de l'Eq. 5. Pour chaque mot et chaque contexte, les racines 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.

4.3 Résultats expérimentaux

Des expériences préliminaires ont été conduites sur cette base de données. Nous nous sommes intéressés au modèle trigram. La capacité du modèle du langage est mesurée en terme de perplexité. Le trigram classique a produit une perplexité de 376.237 sur les données de test avec 17% des mots hors vocabulaire. Le trigram à base de classes morphologiques, utilisant une seule classe pour les racines, a produit une perplexité de 893.127 avec un taux de données

hors vocabulaire de 16.4%. En introduisant deux classes pour les racines, verbes et non-verbes, nous obtenons une perplexité de 877.369 avec le même taux de mots hors vocabulaire. Ces résultats montrent que le modèle à base de classes morphologiques donne des performances médiocres par rapport à un trigram classique. Ceci est intuitivement attendu vues les simplifications effectuées dans notre modèle.

5 Conclusions et Perspectives

Dans ce papier, nous proposons un modèle de langage statistique du type N-grams intégrant des contraintes morphologiques sous la forme de classes. Un cadre théorique complet est présenté. Diverses modèles sont dérivés en effectuant des simplifications plus ou moins importantes.
Ce modèle à base de classes morphologiques a été développé pour la langue Arabe. Il suppose l'existence d'un système d'analyse morphologique. Dans ce travail, des règles d'analyse morphologique empiriquement définies sont utilisées sous la forme d'un transducteur d'analyse. Le logiciel SRILM a été modifié pour intégrer le modèle le plus simplifié. Des tests préliminaires montrent une dégradation de la perplexité par le modèle proposé. Ceci est attendu vu la réduction du nombre de paramètres dans la modélisation. La perte en perplexité est compensée par une réduction de la complexité du modèle.
Le travail sera poursuivi dans diverses directions. D'abord une introduction d'un modèle morphologique plus riche devra être effectuée en incorporant des approches existantes [Beesley 1996][Dichy 1989]. Ensuite, un modèle à base de classes morphologiques moins simplifié que celui utilisé doit être testé.

Remerciements

Ce travail a été partiellement supporté par le projet NEMLAR.

Références

K. BEESLEY, "Arabic Finite-State Morphological Analysis and Generation," *COLING-96*, Vol. 1, pp. 89-94, 1996.

BROWN P. F., DELLA PIETRA V. J., DESOUZA P. V., LAI J. C., AND MERCER R. L., Class-based N-gram models of natural language, *Computational Linguistics*, vol. 18, pp. 467–479, 1992.

CLARKSON P. AND ROSENFELD R., Statistical language modeling using the CMU-Cambridge toolkit, in *Proc*. *EUROSPEEC*H, Rhodes, Greece, Sep. 1997.

DICHY J., Vers un modèle d'analyse automatique du mot graphique non vocalisé en Arabe, Dichy et Hassoun eds. 1989.

MANNING C. AND SCHUTZE H., *Foundations of Statistical Natural Language Processing*, MIT Press. Cambridgw, May 1999.

STOLCKE A., ''SRILM - An Extensible Language Modeling Toolkit,'' in *Proc*. *Intl*. *Conf*. *Spoken Language Processing*, Denver, Colorado, September 2002.