Lhoussain Aouragh & Jelloul Allal
Département de mathématique et informatique,
Faculté des sciences, Université Mohamed premier, Oujda, Maroc,
Abdellah Yousfi
Faculté des sciences juridiques, économiques et sociales, Université Mohamed V- Souissi, Rabat, Maroc

Comparaison entre le modèle de langage p-contexte et le modèle de langage p-grams

Résumé

La modélisation du langage a pour objectif de résumer les connaissances générales liées à un langage naturel. Plusieurs types de modèle de langage ont vu le jour ces dernières années, on cite par exemple le modèle n-grams, le modèle à base d'arbres de décision, etc. Dans ce cadre, nous avons élaboré un modèle de langage appelé modèle p-contexte. Dans cet article nous allons présenter ce modèle et son utilisation, ainsi une comparaison entre les deux modèles p-gram et p-contexte.

1. Introduction

Les modèles de langage sont très répandus et sont nécessaires pour le fonctionnement de plusieurs applications dans le domaine du traitement automatique de la langue, comme
par exemple la traduction automatique, la reconnaissance automatique de la parole continue, etc.
Ces modèles ont pour objectif de résumer les connaissances générales liées à un langage naturel.

Deux types de modèles de langage existent dans la littérature : le modèle à base de connaissances (à base de règles) et le modèle probabiliste. Le premier type nécessite une expertise linguistique pour sa construction et s'appuie généralement sur les grammaires formelles. Cependant ces modèles posent problème dans le cas de leur utilisation dans la reconnaissance de la parole continue, car ils peuvent rejeter des hypothèses de reconnaissance correcte lorsque le locuteur commet des erreurs de grammaire. Le deuxième type, utilise des notions statistiques pour calculer les probabilités de succession des mots.

Plusieurs modèles de langage ont été apparus ces dernières années, le modèle n-gram [1] issu de la théorie de l'information [2], reste toujours un modèle de référence dans plusieurs modèles de langage, comme par exemple les modèles basés sur des arbres de décision [3], les modèles de langage structurés [4], les modèles n-multigrams [5] et les modèles à mémoire cache [6]. L'inconvénient de ces modèles est le nombre énorme de paramètres à estimer, de plus, ce sont des modèles qui demandent un corpus d'apprentissage de taille très grande et bien choisi pour couvrir tous les événements de successions des mots, ce qui n'est pas toujours facile. L'utilisation de ces modèles est resté toujours liée à la parole continue. Dans cet

article, nous avons développé un modèle de langage probabiliste qui est utilisé avec un autre modèle syntaxique pour générer automatiquement une phrase à partir d'un ensemble de mots arabes. Ce modèle est appelé modèle p-contexte. Dans le reste de cet article, on présentera en détail ce modèle ainsi sa comparaison avec le modèle p-gram.

2. Modèle de langage p-gram

Les modèles de langage probabilistes permettent d'attribuer une probabilité à une suite de mots. Ils ont été utilisés pour la première fois dans la reconnaissance automatique de la parole continue. Ces modèles présentent plusieurs avantages par rapport aux modèles à base de connaissances: ils donnent une information quantifiée sur la validité d'une hypothèse contrairement aux modèles à base de connaissances qui retournent des résultats limités de type "acceptation/rejet". L'inconvénient majeur des modèles probabilistes est la quantité importante de données d'apprentissage utilisées pour leur construction.

Les modèles de langage probabiliste sont apparus dans le but de les introduire dans les modèles de la reconnaissance automatique de la parole continue en attribuant une probabilité à une suite de mots. Si on prend une suite de mots $Ph = w_1 w_2 ... w_n$, la probabilité d'avoir cette séquence dans cet ordre est donnée par :

$$\Pr(w_1, w_2, ..., w_N) = \Pr(w_1) \times \prod_{i=2}^{n} \Pr(w_i | w_1, ..., w_{i-1})$$

La construction du modèle probabiliste dans cet état est presque impossible sur le plan expérimental, car d'une part il y a un très grand nombre de probabilités à calculer et d'autre part ceci nécessite un corpus d'apprentissage qui regroupe toutes les phrases possibles pour ce langage, ce qui est impossible.
Pour remédier à ces problèmes, on doit procéder à une approximation de $\Pr(w_i | w_1, ..., w_{i-1})$ en remplaçant l'historique $h_i = w_1, ..., w_{i-1}$ par un historique $F(h_i)$ de taille inférieure à la taille de h_i.
Parmi les approximations les plus utilisées est celle des modèles p-grams. Dans ce cas $\Pr(w_i | h_i)$ ne dépend que des *n-1* derniers mots:

$$\Pr(w_i | h_i) = \Pr(w_i | w_{i-1}, ..., w_{i-n+1})$$

Si le vocabulaire utilisé est de taille assez large, la valeur de *n* ne doit pas dépasser 3.
• Si *n=1*, le modèle est appelé modèle uni-grame. Ce modèle ne prend en compte aucun historique:

$$\mathbf{Pr}(w_i | h_i) = \mathbf{Pr}(w_i)$$

• Si *n=2*, le modèle est appelé modèle bi-grammes. Ce modèle ne prend en compte que le dernier mot:

$$\Pr(w_i | h_i) = \Pr(w_i | w_{i-1})$$

• Si *n=3*, le modèle est appelé modèle tri-grammes. Ce modèle ne prend en compte que les deux derniers mots :

$$\Pr(w_i | h_i) = \Pr(w_i | w_{i-1}, w_{i-2})$$

On a voulu utiliser le modèle n-grams dans la génération automatique des phrases arabes, mais le nombre de paramètres à estimer pour ce modèle est très grand, c'est pourquoi on a élaborer un nouveau modèle qui permet de calculer la probabilité d'un mot sachant son historique sans prendre en compte l'ordre des mot dans cet historique.

3. Modèle de langage p-contexte

Le modèle p-contexte est un modèle qui permet de calculer les probabilités suivantes:

$$\Pr\left(w_{i_j}, w_{i_{j_1}}, ..., w_{i_{j_p}} \text{ se trouvent dans le même contexte }\right)$$

On note par la suite cette probabilité: $\Pr\left(w_{i_j}, w_{i_{j_1}}, ..., w_{i_{j_p}} / contexte\right)$

▪ Pour p = 1: le modèle est appelé modèle bi-contexte et consiste à calculer les probabilités $\Pr\left(w_{i_j}, w_{i_{j_1}} / contexte\right)$.

▪ Pour p = 2: le modèle est appelé modèle tri-contexte et consiste à calculer les probabilités $\Pr\left(w_{i_j}, w_{i_{j_1}}, w_{i_{j_2}} / contexte\right)$.

Ce modèle p-contexte est utilisé avec un autre modèle de syntaxe [Aouragh et al, 2006] pour calculer la probabilité $\Pr\left(w_{i_1}, .., w_{i_n}\right)$. Cette probabilité est sous la forme suivante:

$$\Pr\left(w_{i_1}, .., w_{i_n}\right) = \alpha \times \prod_{j=1}^{n} \prod_{j_1=j+1}^{n} \prod_{j_2=j_1+1}^{n} .... \prod_{j_p=j_{p-1}+1}^{n} \Pr\left(w_{i_j}, w_{i_{j_1}}, ..., w_{i_{j_p}} / contexte\right)$$

α est calculé à l'aide du modèle syntaxique [Aouragh et al, 2006 (b)], [Yousfi et El Jihad, 2005].

3.1. Comparaison entre les deux modèles p-gram et p-contexte

L'inconvénient majeur des modèles p-grams est le nombre de paramètres énormes qui faut estimer, dans le reste de ce paragraphe on donnera une comparaison entre les deux modèles p-contexte et p-grams.

Pour un vocabulaire de taille *K,* on a:

- le nombre de paramètres à estimer pour le modèle bi-gram est: $n^2 + n$
- le nombre de paramètre à estimer pour le modèle bi-contexte est: $\frac{1}{2}n(n-1)$
- Le corpus d'apprentissage nécessaire pour l'estimation des paramètres du modèle p-contexte est de taille inférieur à celui du modèle p-grams.

3.2. Application

Nous avons appliqué ce modèle pour générer automatiquement les phrases arabes. Ce modèle est utilisé avec un autre modèle syntaxique [Aouragh et al, 2006 (a)] pour calculer la probabilité de génération d'une phrase à partir d'une suite de mots.

Nous avons pris le cas p = 1 (modèle bi-contexte). Dans ce cas, la probabilité de générer la phrase $w_{i_1} w_{i_2} ... w_{i_n}$ est:

$$\Pr(w_{i_1},..,w_{i_n}) = \alpha \times \prod_{j=1}^{n} \prod_{k=j+1}^{n} l_{jk}$$

Avec: $\Pr(w_{i_j}, w_{i_k} / contexte) = l_{jk}$

4. Expérimentation

Nous avons construit un corpus d'apprentissage contenant 1449 phrases (de tailles différentes) arabes étiquetées par 186 étiquettes de type syntaxique choisies pour couvrir la majorité des évènements syntaxiques de la langue arabe.

L'évaluation de notre modèle de génération des phrases est réalisée par un programme écrit en langage Perl, contenant deux modules:

- Module d'apprentissage: il permet d'estimer l'ensemble des paramètres de notre modèle.
- Module de génération des phrases: il permet de générer des phrases à partir du vocabulaire de notre système.

Pour évaluer notre modèle nous avons généré toutes les phrases possibles de quatre mots ayant la probabilité de génération non nulle.

Le taux d'erreur utilisé dans notre travail est défini comme étant le pourcentage des phrases fausses générées par rapport à toutes les phrases générées par le système.

Les résultats trouvés sont:

Nombre de phrases générées	7426
Taux d'erreur	61,52%

Références

[8] Aouragh, E., Jelloul, A., Yousfi, A.: 2006, (a) "A Stochastic Language Model for Automatic Generation of Arabic Sentences". Georgian Electronic Scientific Journal: Computer Sciences and Telecommunication |No.**3**(10).

[9] Aouragh, E., Yousfi, A., Jelloul, A.: 2006, (b) INFORSID', Atelier SIA (Systèmes d'Information Arabisés). «La modélisation des niveaux syntaxique et sémantique pour la génération automatique des phrases arabes». Hammamet, Tunisie, 31 mai 2006.

[1] Bahl, L.R., Baker, J.K., Cohen, P.S., Jelinek, F., Lewis, B.L., & Mercer, R.L.: 1978, *Recognition of a continuously read natural corpus*. In : Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing.Tulsa.

[3] Bahl, P., Brown P., De Souza, P. et Mercer, R.: 1990, *A tree-based statistical language model for natural language speech recognition*. Pages 507-514 of: A.WAIBEL et K.-F. LEE (eds), Readings in Speech Recognition. Morgan-Kaufmann.

[4] Chelba, C. et Jelinek, F.: 2000, *Structured language modeling*. Computer, Speech and Language, **14**(4), 283-332.

[5] Deligne, S. et Bimbot, F.: 1995, *Language modeling by variable length sequences: theoretical formulation and evaluation of multigrams*. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing. Détroit, USA.

[2] Jelinek, F.: 1976, *Continuous speech recognition by statistical models*. Proceedings of the IEEE.

[6] Kuhn, R. et De Mori, R.: 1990, *A cache-based natural language method for speech recognition*. IEEE Transactions on Pattern Analysis and Machine Intelligence, **12**(6), 570-582.

[7] Yousfi, A., Jihad, A.: 2005, *Etiquetage morpho-syntaxique*. RECITAL, Durbin, France, 6 10 Juin 2005.