Modèle p-contexte pour la génération automatique des phrases arabes

J. Allal [1], *L.Aouragh*[2], *A.Yousfi*[3]

(1),(2) Faculté des sciences, Dep de Mathématiques, Oujda, Maroc
(3) Institut d'Etudes et de Recherches pour l'Arabisation, Université Med V, Rabat, Maroc
(1)Allal@sciences.univ-oujda.ac.ma
(2) aouragh@hotmail.com
(3) yousfi240ma@yahoo.fr

ABSTRACT

The generation of arabic sentences is a very important operation in the automatic treatment of language. It can be used in several fields : machine translation, continuous speech recognition[1], ...
In this article we developed a stochastic model which measures the probability of generating a arabic sentence from a sequence of words. This model is based on the fact that the sentence consists of two levels: syntactic and semantic, these levels are be treated independently. Each level is characterised by its model.
The estimation of the parameters of these models is calculated on a training corpus (labelled by syntactic tags).

1. INTRODUCTION

Les modèles de langage les plus utilisés dans le domaine de traitement automatique de la parole sont de nature probabilistes. 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 lié à la parole continue. Dans cet article, nous avons développé un modèle stochastique qui permet de calculer la probabilité de générer automatiquement une phrase à partir d'un ensemble de mots dans la langue arabe. Ce modèle combine entre deux niveaux :
• Modèle de langage inspiré du modèle n-gram, que nous avons appelé modèle p-contexte. L'avantage de ce dernier est que le nombre de paramètres à estimer est inférieur à celui de modèle n-gram. De plus, il ne prend pas en compte l'ordre des mots dans la phrase. Nous avons utilisé ce modèle pour modéliser le niveau sémantique.
• Modèle syntaxique : il permet de gérer l'ordre des mots dans les phrases. Il s'appuie sur le calcul d'un chemin optimal d'étiquettes syntaxiques $s^*_{i_1}, \ldots, s^*_{i_n}$ des mots $w_{i_1}, \ldots, w_{i_n}$.

[1]Si on prend une suite d'observations acoustiques $Y_1, \ldots, Y_T$, la résolution de problème de la reconnaissance automatique de la parole continue revient à trouver la suite des mots $w^*_1, \ldots, w^*_n$ telle que: $w^*_1, \ldots, w^*_n$ =

$$\arg \max_{w_1,\ldots,w_n} Pr(Y_1, \ldots, Y_T / w_1, \ldots, w_n) Pr(w_1, \ldots, w_n)$$

2. MODÈLE PROBABILISTE DE GÉNÉRATION DES PHRASES

La phrase peut être vue comme un élément linguistique ayant deux niveaux : niveau syntaxique et niveau sémantique. Pour pouvoir traiter le problème de génération automatique des phrases arabes, nous avons supposé que ces deux niveaux sont indépendants (ce qui nous permet de traiter chaque niveau indépendant de l'autre).
Pour générer une phrase $w_{i_1} w_{i_2} \ldots w_{i_n}$, nous avons modélisé ces deux niveaux par les deux conditions suivantes :
• La première condition (modélisant le niveau sémantique) est que chaque mot w_{i_j} $j \in \{1, 2, \ldots, n\}$ doit apparaître dans un contexte de taille p $(1 \leq p \leq n)$ avec tous les mots restants, c'est-à-dire:

$$Pr(w_{i_j}, w_{i_{j_1}}, \ldots, w_{i_{j_p}} \text{ se trouvent dans le même contexte }) \neq 0$$

Pour tout j et pour tout

$$j_1, j_2, \ldots, j_p \in \{1, \ldots, j-1, j+1, \ldots, n\}$$

On note par la suite cette probabilité :
$Pr(w_{i_j}, w_{i_{j_1}}, \ldots, w_{i_{j_p}}$ / contexte $) \neq 0$
Ceci est équivalant à:

$$\prod_{j=1}^{n} \prod_{j_1=j+1}^{n} \prod_{j_2=j_1+1}^{n} \cdots \prod_{j_p=j_{p-1}+1}^{n} Pr(w_{i_j}, w_{i_{j_1}}, \ldots, w_{i_{j_p}} \text{ / contexte }) \neq 0 \quad (1)$$

Nous avons appelé ce modèle: modèle p-contexte.
- Pour $p = 1$: le modèle est appelé modèle bi-contexte, la formule (1) devient :

$$\prod_{j=1}^{n} \prod_{k=j+1}^{n} Pr(w_{i_j}, w_{i_k} \text{ / contexte }) = \prod_{j=1}^{n} \prod_{k=j+1}^{n} l_{jk} \neq 0$$

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

- Pour $p = 2$: le modèle est appelé modèle tri-contexte, la formule (1) devient :

$$\prod_{j=1}^{n} \prod_{k=j+1}^{n} \prod_{l=k+1}^{n} Pr(w_{i_j}, w_{i_k}, w_{i_l} \text{ / contexte }) \neq 0$$

- Pour $p = n - 1$: la formule (1) devient :

$$Pr(w_{i_j}, w_{i_{j_1}}, \ldots, w_{i_{j_{n-1}}} \text{ / contexte }) \neq 0$$

• La deuxième condition (modélisant le niveau syntaxique) permet de vérifier si l'ordre des mots $w_{i_1}, w_{i_2}, \ldots, w_{i_n}$ est juste ou non. Ceci est réalisée en se basant sur des connaissances grammaticales des mots $w_{i_1}, w_{i_2}, \ldots, w_{i_n}$.
Nous avons modélisé cette condition par l'existence d'un chemin optimal $s^*_{i_1}, s^*_{i_2}, \ldots, s^*_{i_n}$ d'étiquettes de types syntaxiques des mots $w_{i_1}, w_{i_2}, \ldots, w_{i_n}$, tel que :

$$Pr(w_{i_1}, \ldots, w_{i_n}, s^*_{i_1}, \ldots, s^*_{i_n}) \neq 0 \qquad (2)$$

La probabilité de générer la phrase $w_{i_1} w_{i_2} \ldots w_{i_n}$ est le produit entre les deux probabilités (1) et (2) (les deux niveaux sémantique et syntaxique sont supposés indépendants):

$$Pr(w_{i_1}, \ldots, w_{i_n}) = \beta^*_{d_n} \prod_{j=1}^{n} \prod_{j_1=j+1}^{n} \prod_{j_2=j_1+1}^{n} \ldots$$
$$\prod_{j_p=j_{p-1}+1}^{n} Pr(w_{i_j}, w_{i_{j_1}}, \ldots, w_{i_{j_p}} \text{ / contexte })$$
$$\times Pr(w_{i_1}, \ldots, w_{i_n}, s^*_{d_1}, \ldots, s^*_{d_n}) \qquad (3)$$

où: $\beta^*_{d_n} = Pr(s^*_{d_n}$ soit état finale)

REMARQUE

Nous avons ajouté la probabilité de l'état final pour éliminer la génération des phrases non complètes. Si on prend par exemple la phrase "دخل الولد الَى", la probabilité de génération de cette phrase sans prendre en compte $\beta^*_{d_n}$ est non nulle car dans le corpus d'apprentissage, nous avons la phrase "دخل الولد الَى المدرسة".

3. APPLICATION

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

$$Pr(w_{i_1}, \ldots, w_{i_n}) = \beta^*_{d_n} Pr(w_{i_1}, \ldots, w_{i_n},$$
$$s^*_{d_1}, \ldots, s^*_{d_n}) \prod_{j=1}^{n} \prod_{k=j+1}^{n} l_{jk} \qquad (4)$$

$s^*_{d_1}, s^*_{d_2} \ldots, s^*_{d_n}$: le chemin optimal d'étiquettes syntaxiques associé à la phrase $w_{i_1} w_{i_2} \ldots w_{i_n}$, il est donné par:

$$s^*_{d_1}, s^*_{d_2}, \ldots, s^*_{d_n} =$$
$$\arg \max_{s_{j_1}, \ldots, s_{j_n}} Pr(w_{i_1}, \ldots, w_{i_n}, s^*_{j_1}, \ldots, s^*_{j_n}) \qquad (5)$$

Nous avons utilisé les modèles de Markov cachés [7] pour résoudre le problème (5). On suppose que le double processus (X_t, Y_t) est un modèle de Markov caché (MMC) d'ordre 1 vérifiant les hypothèses suivantes :

• $X_t = s_i$: est une chaîne de Markov d'ordre 1 à valeurs dans un ensemble d'étiquettes syntaxiques $E = \{s_1, \ldots, s_N\}$, X_t vérifie :

- $Pr(X_{t+1} = s_j / X_1 = s_{i_1}, \ldots, X_t = s_k) = Pr(X_{t+1} = s_j / X_t = s_k) = a_{kj}$

- $Pr(X_1 = s_i) = \pi_i \quad i \in \{1, \ldots, N\}$.

• $Y_t = w_i$ est un processus à valeurs dans un ensemble de mots $V = \{w_1, \ldots, w_M\}$ représentant le vocabulaire de notre système, Y_t vérifie :
$Pr(Y_t = w_t / X_1 = s_{i_1}, ..., X_t = s_j, Y_{t-1} = w_{t-1}, ..., Y_1 = w_1) =$
$= Pr(Y_t = w_t / X_t = s_j) = b_j(w_t) = b_{jt}$
b_{jt} : la probabilité que le mot w_t a l'étiquette s_j.

REMARQUE

Notre modèle de génération des phrases est défini entièrement par un vecteur de paramètres noté
$\Theta = (\Pi, \beta, A, B, L)$:
- $\Pi = \{\pi_1, \ldots, \pi_N\}$ l'ensemble des probabilités des étiquettes initiales.
- $\beta = \{\beta_1, \ldots, \beta_N\}$ l'ensemble des probabilités des étiquettes finales.
- $A = (a_{ij})_{1 \leq i \leq N, 1 \leq j \leq N}$ la matrice des probabilités de transition entre les étiquettes.
- $B = (b_{it})_{1 \leq i \leq N, 1 \leq t \leq M}$ la matrice des probabilités que le mot w_t a l'étiquette s_i.
- $L = (l_{ij})_{1 \leq i \leq M, 1 \leq j \leq M}$ la matrice des probabilités d'apparition d'un mot w_i dans le même contexte que w_j.

3.1. Calcul du chemin optimal

Pour le calcul du chemin optimal, on utilise l'algorithme de Viterbi.
On définit $\delta_t(s_k)$ par la probabilité du meilleur chemin partiel aboutissant à l'étiquette s_k à l'instant t.

$$\delta_t(s_k) = \max_{s_{d_1}, \ldots, s_{d_t}} Pr(w_{i_1}, \ldots, w_{i_t}, s_{d_1}, \ldots, s_{d_{t-1}}, s_k)$$

La règle de Bayes nous donne la formule récurrente suivante :

$$\delta_t(s_k) = \max_{s_j} [\delta_{t-1}(s_j) a_{jk} b_k(w_t)] \quad (6)$$

$\forall\, t \in \{i_1, \ldots, i_M\}\ et\ \forall\, k \in \{j_1, \ldots, j_N\}$
Le chemin optimal est obtenu à l'aide d'un calcul récursive sur la formule (6).

3.2. Estimation des paramètres du modèle

En général, trois méthodes d'estimation de ces paramètres peuvent être utilisées : l'estimation de Maximum de vraisemblance [8], l'estimation par Maximum à Posteriori [9] et l'estimation par maximum d'information mutuel [3].
Dans notre cas, nous avons utilisé le maximum de vraisemblance. Pour un ensemble d'apprentissage $R = \{ph_1, \ldots, ph_K\}$ constitué par un ensemble de phrases arabes étiquetées par un ensemble d'étiquettes syntaxiques $E = \{s_1, \ldots, s_N\}$, l'estimation de Θ est donnée par :

$$\Theta^* = arg[\max_{\Theta} \prod_{i=1}^{K} Pr_{\Theta}(ph_i)] \quad (7)$$

La résolution de ce problème donne les formules d'estimation suivantes :

- $\pi_i = \frac{\sum_{j=1}^{k} \delta(s_i \text{ est état initial dans } pk_j)}{K}$
- $\beta_i = \frac{\sum_{j=1}^{k} \delta(s_i \text{ est état final dans } pk_j)}{K}$

- $a_{ij} = \frac{\sum_{l=1}^{k} F_l(s_i s_j)}{\sum_{l=1}^{k} F_l(s_i)}$

- $b_{ij} = \frac{\sum_{l=1}^{k} F_l(w_j \text{ a l'étiquette } s_i)}{\sum_{l=1}^{k} F_l(s_i)}$

- $l_{ij} = \frac{\sum_{l=1}^{k} F_l(w_i \text{ et } w_j)}{K}$

où :
$\delta(s_i$ est état initial dans $ph_j)$=
$\begin{cases} 1 & \text{si } s_i \text{ est un état initial dans } ph_j \\ 0 & \text{sinon.} \end{cases}$
$F_i(\sigma)$ est le nombre de fois où σ est dans la phrase ph_i

3.3. Expérimentation

Données d'apprentissages Nous avons construit un corpus d'apprentissage contenant 1449 phrases (de tailles déférentes) arabes étiquetées par 186 étiquettes de type syntaxique choisies pour couvrir un peu tous les 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.

Résultats 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.
Le taux d'erreur exacte sur ces phrases est le suivant :

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

On remarque que le taux d'erreur dans ce cas est très élevé. La plupart de ces erreurs proviennent essentiellement du niveau syntaxique (la structure syntaxique de plusieurs phrases générées est extraite exactement de structures de phrases de longueur différent de quatre mots).
Pour remédier à ce problème, nous avons élaboré deux approches :
• la première approche utilise seulement les phrases de quatre mots pour faire l'apprentissage du modèle de génération, les résultats obtenus pour cette approche sont donnés par :

Nombre de phrases générées	592
Taux d'erreur	7,43%

Le taux d'erreur est diminué considérablement par rapport au premier cas, mais le nombre de phrases générées est très réduit, il représente seulement 8% des phrases générées dans le premier cas.
• Pour la deuxième approche, la procédure d'apprentissage du MMC est faite seulement sur les phrases de quatre mots, tandis que l'apprentissage du modèle bi-contexte est fait sur toutes les phrases du corpus d'apprentissage. Les résultats obtenus sont:

Nombre de phrases générées	1193
Taux d'erreur	29.08%

On remarque que le nombre de phrases générées est augmenté deux fois par rapport au deuxième cas. Le taux d'erreur a été réduit (par rapport au premier cas) de 52.73%.
L'analyse de tous ces résultats, montre que la plupart des erreurs proviennent essentiellement des points suivants :
• L'ordre du modèle de langage que nous avons utilisé est un, plusieurs erreurs vont être éliminés pour l'ordre deux et trois.
• Les étiquettes ne sont pas bien spécifiques, ceci donne des phrases justes au niveau syntaxique et fausses au niveau sémantique.

سافر أبي في الطائرة
دخلت زينب إلى المنزل
جلس محمد قرب المدرسة
حضر أصدقائي إلى السينما
وصل أبي إلى المنزل
عاد محمد من المدرسة
صلى الولد مع الجماعة
انتظر لحظة من المريض

Figure 1: Un extrait des phrases générées par notre modèle.

4. CONCLUSIONS ET PERSPECTIVES

Les résultats obtenus sont en générale encourageants (car les travaux dans ce sens sont rares), pour diminuer l'erreur de génération des phrases, comme perspectives, nous allons augmenter l'ordre du modèle p-contexte. Ceci augmentera le nombre de paramètres à estimer, l'utilisation da la notion de classes et leur choix, dans ce cas, deviendra nécessaire et important, le préférable est d'utiliser des classes combinant entre les deux niveaux sémantique et syntaxique pour les mots.
De même, il est très utile de réaliser l'apprentissage du modèle p-contexte indépendamment du modèle MMC et sur un corpus d'apprentissage plus large (car dans ce cas, nous ne somme pas obligé à étiqueter manuellement ce corpus).
Théoriquement, les paramètres du modèle de génération doivent converger vers des valeurs constantes, un apprentissage sur un corpus de taille très importante approchera ces paramètres de ces valeurs, ceci permettera de déduire les probabilité (taux) d'utilisation des phrases dans la langue arabe.

BIBLIOGRAPHIE

[1] BAHL L.R., BAKER J.K., COHEN P.S., JELINEK F., LEWIS B.L., et MERCER R.L., *Recognition of a continuously read natural corpus*. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing.Tulsa. 1978

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

[3] BAHL P., BROWN P., DE SOUZA P. et MERCER R., *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, 1990.

[4] CHELBA C. et JELINEK F., *Structured language*

modeling. Computer, Speech and Language, 14(4), 283-332, 2000.

[5] DELIGNE S. et BIMBOT F., *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, 1995.

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

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

[8] FEDERICO M. et DE MORI R., *Language Modelling*. Chap. 7, pages 204-210 of : R. DE MORI (ed), Spoken Dialogue with Computers. Academic Press, 1998b.

[9] Yannick E., *Intégration de sources de connaissances pour la modélisation stochastique du langage appliquée à la parole continue dans un contexte de dialogue oral homme-machine*. Thèse de Doctorat, Université d'Avignon, 2002.