Lhoussain AOURAGH (1) **, Jelloul ALLAL** (2) **, Abdellah YOUSFI** (3)
(1) et (2) Département de mathématiques et informatique, faculté des sciences, Université Mohamed premier, Oujda, Maroc,
(3) Institut d'Etudes et de Recherches pour l'Arabisation, Rabat - Maroc
(1)aouragh@hotmail.com
(2)Allal@sciences.univ-oujda.ac.ma
(3)yousfi240ma@yahoo.fr

Modèle p-contexte pour la génération automatique des phrases arabe

Résumé

La modélisation du langage a pour objectif de résumer les connaissances générales liées à un langage naturel. Dans ce cadre, la génération des phrases est une opération très importante dans le traitement automatique de langage. Elle peut être utilisée dans plusieurs domaines par exemple la traduction automatique, la reconnaissance de la parole continue, etc.

Dans cet article, nous avons élaboré un modèle stochastique qui permet de mesurer la probabilité de générer une phrase dans la langue arabe à partir d'un ensemble de mots. Ce modèle s'appuie sur le fait que la phrase est constituée de deux niveaux indépendants syntaxique et sémantique, ce qui nous a permis de caractériser chaque niveau par son propre modèle.

I . Introduction :

Les modèles de langage les plus utilisés dans le domaine de traitement automatique de la parole sont de nature probabiliste. 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 : syntaxique et sémantique.

II . Modèle stochastique de génération des phrases arabe

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 indépendamment de l'autre).

La génération d'une phrase $w_{i_1} w_{i_2} \ldots . w_{i_n}$ suppose les deux conditions suivantes :

i) Chaque mot w_{i_j} $j \in \{1,...,n\}$ doit apparaître dans un contexte de taille p $(1 \le p \le n)$ avec tous les mots restants.

ii) Il existe un chemin optimal $s^*_{i_1}, s^*_{i_2}, ..., s^*_{i_n}$ d'étiquettes de types syntaxiques des mots $w_{i_1}, w_{i_2}, ..., w_{i_n}$ tel que :

$$\Pr\left(w_{i_1}, ..., w_{i_n}, s^*_{i_1}, ..., s^*_{i_n}\right) \neq 0$$

Conséquences

1) La condition (i) implique que pour tout j et pour tout $j_1, ..., j_p \in \{1, .., j-1, j+1, .., n\}$ on a :

$$\Pr\left(w_{i_j}, w_{i_{j_1}}, ..., w_{i_{j_p}} \; se\,trouvent\,dans\,le\,même\,contexte\right) \neq 0$$

On note par la suite cette probabilité : $\Pr\left(w_{i_j}, w_{i_{j_1}}, ..., w_{i_{j_p}} / contexte\right) \neq 0$

Ceci est équivalant à :

$$\prod_{j=1}^{n}\prod_{j_1=j+1}^{n}\prod_{j_2=j_1+1}^{n}....\prod_{j_p=j_{p-1}+1}^{n}\Pr\left(w_{i_j},w_{i_{j_1}},...,w_{i_{j_p}}/contexte\right)\neq 0 \qquad (1)$$

Nous avons appelé ce modèle : modèle p-contexte.

- Pour p = 1 : le modèle est appelé modèle bi-contexte.
- Pour p = 2 : le modèle est appelé modèle tri-contexte.

2) La condition (ii) permet de vérifier si l'ordre des mots $w_{i_1},w_{i_2},...,w_{i_n}$ est juste ou non. Ceci est réalisé en se basant sur des connaissances grammaticales des mots $w_{i_1},w_{i_2},...,w_{i_n}$.

Nous avons modélisé cette condition par l'existence d'un chemin optimal $s_{i_1}^*,s_{i_2}^*,...,s_{i_n}^*$ d'étiquettes de types syntaxiques des mots $w_{i_1},w_{i_2},...,w_{i_n}$ tel que :

$$\Pr\left(w_{i_1},...,w_{i_n},s_{i_1}^*,...,s_{i_n}^*\right)\neq 0 \qquad (2)$$

3) La synthèse des deux conditions donne la probabilité de générer la phrase $w_{i_1}w_{i_2}...w_{i_n}$ comme étant le produit entre les deux probabilités (1) et (2) (les deux niveaux sémantique et syntaxique sont supposés indépendants) :

$$\begin{aligned}\Pr\left(w_{i_1},..,w_{i_n}\right)=\beta_{d_n}^*\times&\prod_{j=1}^{n}\prod_{j_1=j+1}^{n}\prod_{j_2=j_1+1}^{n}....\prod_{j_p=j_{p-1}+1}^{n}\Pr\left(w_{i_j},w_{i_{j_1}},...,w_{i_{j_p}}/contexte\right)\\&\times\Pr\left(w_{i_1},...,w_{i_n},s_{d_1}^*,...,s_{d_n}^*\right)\end{aligned} \qquad (3)$$

où : $\beta_{d_n}^*=\Pr\left(s_{d_n}\text{ soit état finale}\right)$

III . Application

Comme application de ce modèle, 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 :

$$\Pr\left(w_{i_1},..,w_{i_n}\right)=\beta_{d_n}^*\times\Pr\left(w_{i_1},...,w_{i_n},s_{d_1}^*,...,s_{d_n}^*\right)\times\prod_{j=1}^{n}\prod_{k=j+1}^{n}l_{jk} \qquad (4)$$

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

$$s_{d_1}^*,s_{d_2}^*,...,s_{d_n}^*=\arg\max_{s_{j1},..,s_{jn}}\Pr\left(w_{i_1},..,w_{i_n},s_{j_1},..,s_{j_n}\right) \qquad (5)$$

Nous avons utilisé les modèles de Markov cachés [7] caractérisé par l'ensemble de paramètres $\left(\Pi,\beta,A,B\right)$ pour résoudre le problème (5).

III.1 Remarque

Notre modèle de génération des phrases est défini entièrement par un vecteur de paramètres noté $\Theta=\left(\Pi,\beta,A,B,L\right)$.

Pour estimer ces paramètres, nous avons utilisé l'estimation par maximum de vraisemblance.

III.2 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.

III.3 3.3.2 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.

Les expériences sont faites selon trois approches :

- L'apprentissage de tous les paramètres est fait sur tout le corpus.
- L'apprentissage du modèle de génération utilise seulement les phrases de quatre mots.
- L'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.

Les résultats trouvés sont :

	Approche 1	Approche 2	Approche 3
Nombre de phrases générées	7426	592	1193
Taux d'erreur	61,52%	7,43%	29.08%

IV . Références

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.