K.Meftouh[(1)] - K.Smaïli[(2)] - M.T.Laskri[(1)]
[(1)] Université Badji Mokhtar de Annaba
Département d'informatique, équipe LRI/GRIA
BP 12 23000, Annaba, Algérie.
karima.mef@voila.fr, laskri@univ-annaba.org
[(2)] INRIA-LORIA, équipe Parole
BP 101 54602 Villers Les Nancy, France
smaili@loria.fr

Modélisation statistique de la langue Arabe

1. Introduction

Le rôle d'un modèle de langage est de modéliser les différents événements langagiers de la langue qui sont modélisés à partir de corpus. A chaque événement une probabilité d'occurrence est estimée. Un modèle de langage statistique estime la probabilité $P(w_1^n)$ d'une suite de mots $w_1^n = w_1 w_2 ... w_n$. Cette dernière est égale au produit des probabilités des suites de mots qui la composent [7]:

$$P\left(w_1^n\right) = \prod_{i=1}^{n} P(w_i / w_1, w_2, ..., w_{i-1}) \quad (1)$$

Où $P\left(w_i / w_1, w_2, ..., w_{i-1}\right)$ est la probabilité d'apparition du mot w_i sachant la suite de mots $w_1, w_2, ..., w_{i-1}$ précèdent directement w_i. Cette suite de mots est appelée l'historique de w_i.

Dans la pratique il est difficile d'estimer correctement ces probabilités. En effet, lorsque *n* est important, il est impossible de trouver des historiques de taille *n-1* suffisamment fréquents pour les estimer correctement [7]. C'est pourquoi la formule (1) est généralement approximée comme suit :

$$P\left(w_1^n\right) = \prod_{i=1}^{n} P(w_i / w_{i-k+1}, ..., w_{i-1}) \quad (2)$$

Dans la littérature *k* prend la valeur 2, 3 ou 4 correspondant respectivement aux modèles 2-grammes, 3-grammes ou 4-grammes.

Plus formellement, un modèle de langage peut être défini par un triplet (C, V, P) où C désigne le corpus d'apprentissage, V une liste de mots constituant le vocabulaire de l'application et P une distribution de probabilités conditionnelles [8]. Dans cet article, nous nous intéressons à la modélisation statistique de la langue Arabe. L'objectif étant d'étudier son adéquation à être modélisé par des n-grammes classiques. Nous avons donc développé des modèles n-grammes d'ordre 2, 3 et 4 en utilisant le CMU-Toolkit [12].

2. La nécessité du lissage

Le lissage est l'opération permettant d'ajuster l'estimation au maximum de vraisemblance des probabilités afin d'obtenir des mesures de probabilités plus précises.Ces techniques permettent en général de modifier les probabilités soit de certains événements ou de tous les événements. Les techniques de lissage ont comme principe d'altérer les probabilités initiales calculées par un simple comptage normalisé des occurrences des événements [1]. Cette altération est plus que nécessaire, sinon certains événement non rencontrés en phase d'apprentissage se verront affecter une probabilité nulle. Cela condamnerait à jamais en phase de test une phrase dont un de ses composants a été absent de l'apprentissage.

Plusieurs mots de la langue Arabe respectent une topologie morphologique précise. Pour prendre en compte cette spécificité, nous avons menés plusieurs expérimentations pour trouver le meilleur type de modèle de langage. Nous montrerons dans la suite que les meilleurs résultats sont ceux pour lesquels l'apprentissage a été effectué sur des corpus où chaque mot a été segmenté partiellement en ses formes infra-lexicales.

3. Expérimentations et évaluation

Les corpus d'apprentissage et de test sont extraits du quotidien Algérien « Al-khabar ». Nous avons utilisé 79500 mots pour l'apprentissage et 5040 mots pour le test. Le vocabulaire est constitué des 2000 mots (incluant le mot virtuel inconnu noté UNK) les plus fréquents. Différentes méthodes de lissage ont été utilisées : linéaire,

Good Turing et Witten Bell (voir [11] pour une étude détaillée des différentes méthodes de lissage). Dans tous les résultats présentés ci-dessous, les modèles sont évalués en terme de perplexité (P) et d'entropie (E).

I.1 Modèles de langage appris sur des corpus bruts

Le premier modèle développé est calculé à partir de n-grammes en ne procédant à aucun prétraitement des textes. Des modèles d'ordre 2, 3 et 4 ont été ainsi appris. Les tableaux Tab 1 et Tab 2 donnent les valeurs de perplexité obtenues avec et sans UNKs. Le pourcentage de mots hors vocabulaire est de 30.19%.

n	**Good-Turing**		**Witten-bell**		**Linéaire**	
	P	E	P	E	P	**E**
2	289.10	8.18	267.86	8.07	309.29	8.27
3	292.36	8.19	278.87	8.12	321.50	8.33
4	307.51	8.26	311.97	8.29	335.14	8.39

Tab1: Perplexité des modèles d'ordre 2, 3 et 4 calculées sans UNKs.

Tab2 : Perplexité des modèles d'ordre 2, 3 et 4 calculées avec UNKs.

n	**Good-Turing**		**Witten-bell**		**Linéaire**	
	P	E	P	E	P	**E**
2	76.66	6.26	76.03	6.25	82.92	6.37
3	81.55	6.35	81.18	6.35	92.09	6.52
4	88.07	6.46	89.25	6.48	97.67	6.61

I.2 Modèles de langage appris sur des corpus de morphèmes

Un mot arabe est constitué d'une séquence de morphèmes selon le schéma préfixes*-radical-suffixes* (* désigne zéro ou plusieurs occurrences de morphèmes) [3]. Pour améliorer les performances des modèles n-grammes, nous avons décidé de procéder à un premier prétraitement qui consiste à segmenter partiellement les mots constituants les corpus d'apprentissage et de test, et de calculer ainsi des modèles à base de morphèmes plutôt que de mots. Par segmentation partielle nous entendons séparer le(s) préfixe(s) du reste du mot. Des exemples de mots arabes et leurs segmentations sont donnés dans la table Tab 3. La table Tab 4 liste l'ensemble des préfixes auxquels nous nous sommes intéressés.

mot	préfixes	Reste du mot
الولايات	الـ	ولآيات
وليقوم	لـ, و	يقوم

Tab3: exemples de mots arabes et leur segmentation

Préfixes	
و	كـ
بـ	فـ
لـ	الـ

Tab 4 : Ensemble de préfixes

Le deuxième prétraitement concerne les noms de villes composés tels que: سوق أهراس

Ces noms doivent être considérés comme des entités du vocabulaire à part entière [9]. Pour cela, tous les noms de villes composés ont été modifiés de façon à ne former qu'un seul mot: سوق_أهراس. Ainsi, nous obtenons un corpus d'apprentissage de 110000 mots et 6960 mots pour le test. Les valeurs de perplexité données dans les tableaux Tab 5 et Tab 6 montrent que les modèles appris sur de tels corpus sont nettement plus performants que ceux calculés en 1. Le taux de mots or vocabulaire est de 15.92% au lieu de 30.19%. Ces résultats prouvent ainsi que la segmentation du texte est une étape indispensable dans la modélisation de la langue Arabe.

n	**Good-Turing**	**Witten-Bell**	**Linéaire**

	P	E	P	E	P	**E**
2	87.89	6.46	86.44	6.43	94.25	6.56
3	68.42	6.10	65.44	6.03	75.50	6.24
4	69.82	6.13	66.70	6.06	76.08	6.25

Tab 5 : Perplexité (sans UNKs) des modèles appris sur des corpus partiellement segmentés.

n	**Good-Turing**		**Witten-Bell**		**Linéaire**	
	P	E	P	E	P	**E**
2	57.63	5.85	57.66	5.85	61.91	5.95
3	47.27	5.56	46.17	5.53	52.81	5.72
4	48.72	5.61	47.11	5.56	53.96	5.75

Tab 6 : Perplexité (avec UNKs) des modèles appris sur des corpus partiellement segmentés.

4. Conclusion

Dans ce travail, nous nous sommes intéressés à la modélisation statistique de la langue Arabe. Des modèles n-grammes ont été calculés. Les expérimentations effectuées ont montré que les modèles appris sur des corpus, où les mots ont été segmentés, sont nettement plus performants (15.92% de mots hors vocabulaire au lieu de 30.19%). Ces résultats prouvent ainsi que la segmentation du texte est indispensable pour la modélisation statistique de l'Arabe. Les résultats en terme de perplexité ont montré également la suprématie des modèles à base de morphèmes. Notre objectif est de combiner les morphèmes et les mots au sein d'une structure plus complexe qui est supportée par les réseaux Bayésiens.

5. Références bibliographiques

1. J.P. Haton, C. Cerisara, D. Fohr et K. Smaïli « Reconnaissance automatique de la parole – Du signal à son interprétation », A paraître chez Dunod à partir du 25 mai 2006.

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

3. Y.Lee, K.Papineni, S.Roukos, O.Emam, H.Hassan, *Language model based Arabic word segmentation*. In proc. of the 41st annual meeting of the association for computational linguistics, July 2003, 399-406.

4. Andreas Stolcke, SRILM, An extensible language modeling toolkit. In proc.Intl.conf spoken language processing, Denver, Colorado, September 2002.

5. K.Kirchoff et al., *Novel approaches to Arabic speech recognition*. Final report from the 2002 Johns-Hopkins summer workshop. John-Hopkins University, Tech.report,2002.

6. Kareem Darwish, Building a shallow Arabic morphological analyzer in one day. In proceedings of the ACL workshop on computational approaches to Semitic languages, Philadelphia, PA, 2002.

7. Joshua T.Goodman, A bit of progress in language modeling, extended version. Technical report MSR-TR-2001-72, August 2001.

8. K.Smaïli, Les modèles de langage statistiques: de la reconnaissance à la traduction. HDR, Université Nancy 2, Nancy, France 2001.

9. I.Zitouni, K.Smaïli, vers une meilleure modélisation du langage: la prise en compte des séquences dans les modèles statistiques. Journées d'études sur la parole, Aussois, France 2000, 293-296.

10. Egyptian Demographic center, 2000.

11. http://www.frcu.eun.eg/www/homepage/cdc/cdc.htm

12. Stanley F.Chen, J.Goodman, An empirical study of smoothing techniques for language modeling. Tech. report TR-10-98, Computer science group, Harvard University, Cambridge, Massachusetts, August 1998.

13. P.Clarkson, R.Rosenfeld, Statistical language modeling using the CMU-Cambridge Toolkit. IN proc.Eurospeech, Rhodes, Greece, September 1997.