Arabic statistical modeling based on morphology (Modélisation statistiques basée sur la morphologie pour la langue arabe)

Ali Sadiqui, Noureddine Chenfour
Faculté des Sciences Dhar El Mehraz,
Université Sidi Mohamed Ben Abdellah de Fès, Morocco

ABSTRACT: In this work we submit the results obtained for the building of a statistical model of the Arabic language, adopting for a word the prefix*-stem-suffix structure based on the lattice. That solution allowed us to keep all the possibilities of word segmentation, which is one of the issues we have met when building the aforementioned model. The language has been evaluated from a corpus made up of 100K words and has been tested on a corpus of 7K words. The results and the analysis are submitted in this document.
KEYWORDS: Automatic Speech Recognition, Language Model, Arabic Language, SRILM.

Introduction

Dans la langue Arabe, un mot se compose d'un radical ou une tige entourée par des préfixes et/ou des suffixes. Les affixes signalent des catégories grammaticales telles que la personne, le nombre et le genre. Le radical lui aussi peut, plus loin, être décomposé en *racine* (souvent ordre de trois consonnes) et un modèle des voyelles et, probablement, des consonnes additionnelles. Un mot arabe peut être segmenté donc à un ordre de morphèmes ou items lexicaux (*tokens*) respectant le modèle suivant *prefix-stem-suffix*.

Par segmentation, nous entendons ici la segmentation lexicale ou itémisation ((*tokenization* ou *word segmentation*) qui consiste à segmenter un texte en mots-formes ou items lexicaux. C'est une opération consistant à

structurer le texte en passant d'un ensemble continu de caractères à une suite discrète d'items lexicaux.

La réalisation des modèles statistiques de langage basé sur des classes morphologiques prend un grand intérêt, car leur efficacité a été prouvée. Ces modèles ont l'avantage de réduire le nombre de mots qui n'ont pas été encore vus (OOV), et ils nécessitent généralement un espace mémoire réduit.

Cette étude, qui s'ajoute à d'autres déjà effectuées dans ce domaine, a pour but la réalisation d'un modèle statistique de langage basé sur des classes morphologiques, en s'appuyant sur l'analyseur morphologique AraMorph et sur l'outil lattice-tool de SRILM.

Dans la première section de cet article nous commencerons par définir les modèles N-gramme pour passer, dans la deuxième section, à présenter la problématique rencontrée au niveau de la langue Arabe pour la réalisation de ce modèle, présenter la solution adoptée et les résultats obtenues, avant de conclure.

1 Les modèles de N-gramme

Les modèles de langue constituent une des composantes clés dans les systèmes modernes de reconnaissance de la parole. Son but est de déterminer la probabilité d'un ordre de mot, et d'essayer de prévoir le prochain mot dans un ordre de mots.

Ce genre de modèle est utile dans une grande variété de secteurs de recherche à savoir : la reconnaissance de la parole, la reconnaissance optique des caractères, la traduction automatique, …

Les modèles de langue les plus couramment employés sont des modèles de n-gramme.

Cette modélisation correspond en fait à un modèle de Markov d'ordre n, où seules les n dernières observations sont utilisés pour la prédiction de la lettre suivante. Ainsi un bigramme est un modèle de Markov d'ordre 2.

Dans les modèles N-gramme, un mot est estimé selon les n mots précédents.

Cette probabilité est décomposée comme suit :

$$p(s) = p(w1)p(w2|w1)p(w3|w1w2)\ldots p(w1|w1\ldots wi\text{-}1) = \prod_{i=1}^{l} p(wi / w1...wi-1)$$

Pour le de bigramme :

$$p(s) = \prod_{i=1}^{l} p(wi / w1...wi-1) \approx \prod_{i=1}^{l} p(wi / wi-1)$$

Si on introduit la décomposition par classe, cette probabilité devient alors :

$$p(w \mid wi-1) \approx p(wt \mid C(w))p(C(w) \mid C(w-1))$$

Les classes utilisées en général sont des classes à base syntaxique, sémantique ou morphologique ou même des classes déterminées automatiquement. Dans notre travail nous nous sommes intéressé aux classes morphologiques.

Il faut aussi noter qu'à la différence de la linguistique, la structure grammaticale est non pertinente dans les modèles de langage, et n'importe quelle combinaison de mots, même si elle est insensée, on lui assigne souvent une probabilité proche-zéro.

2 Problématique de la langue Arabe

Dans les textes Arabes non vocalisés, la complication qui se présente, est que l'analyse morphologique d'un mot, se basant sur les lexèmes, peut aboutir à plusieurs solutions.

Prenons, comme exemple, le mot « أمسك », ce mot peut présenter trois homonymes et selon le contexte, on peut en tirer trois analyses morphologiques :

1- le verbe « أمسك (attraper/maintenir) » qui n'a ni préfixe ni suffixe
2- le mot « أمس (hier) » avec préfixe « ك »
3- le verbe «مس / (toucher) » avec préfixe « ك » et le suffixe « أ »
Ce qui donne :

Solution	Décomposition
1	أمسك (attraper/maintenir)
2	أ + مس + ك (il vous a touché ?)
3	أمس + ك (votre veillé)

Tableau 1. Les différentes segmentations de mot « أمسك »

Voila un deuxième exemple :

L'analyse morphologique du mot « فهم » peut aboutir à deux solutions

Solution	Décomposition
1	فهم (comprendre)
2	هم+ ف (et ils sont)

Tableau 2. Les différentes segmentations de mot « فهم »

Une seule de ces formes est juste dans la phrase en dépit de leur écriture identique dans un texte non vocalisé. Le choix de la solution exacte dépend de l'ensemble des mots qui l'entoure, et en l'absence de toute base de connaissance préalable il est difficile qu'un système puisse y parvenir.

N° de test	nombre de Mots	nombre de mots qui présentent plusieurs possibilités
Corpus 1 (30K)	29 239	3803 (13%)
Corpus 2 (50K)	51369	7867 (15%)
Corpus 3 (90K)	87626	14413 (16%)
Corpus 4 (150K)	149625	44305 (29%)

Tableau 3. Tableau représentant le nombre de mots qui présentent plusieurs possibilités sur différents corpus

إن + # إعاد + ة ال# نظر + # في + # مراكز + {# تكوين + , ت# كوي + ن
}ال# معلم + ين وال# معلم + ات بال# ضرور + ة ي# ستدعي + ال# توفر + # على
+ { ت# صور +, # تصور + }و# أي + {ت# صور + ,# تصور + }# {حالي +, # حال +
ي}# ل# هاته + ال# مراكز + {ل# كي +, # لك + ي} ي# تمكن + # من + # ملامس + ة
حقيق + ة ال# منظوم + ة ال# تربوي + ة ال# حالي + ة # في + # {أفق +,أ# فق +
}{ ت# حس + ين, # تحسين +}# أدائ + ها ومردو # دي + تها ي# جب + # علي + ه ال#
ارتكاز + # على + ال# ميثاق + ال# وطني + لل# تربي + ة وال# تكوين + و#
{ذلك + ,و# ذل + ك}# لأن + # هذا + ال# ميثاق + {أ# صبح + , # أصبح + }ي#
شكل + ال# مرجعي + ة ال# تشريعي + ة

Figure 1. Exemple de texte où les mots sont segmentés suivant le modèle prefix-stem-suffix. Les mots entre le "{}" présentent d'autres possibilités de segmentation

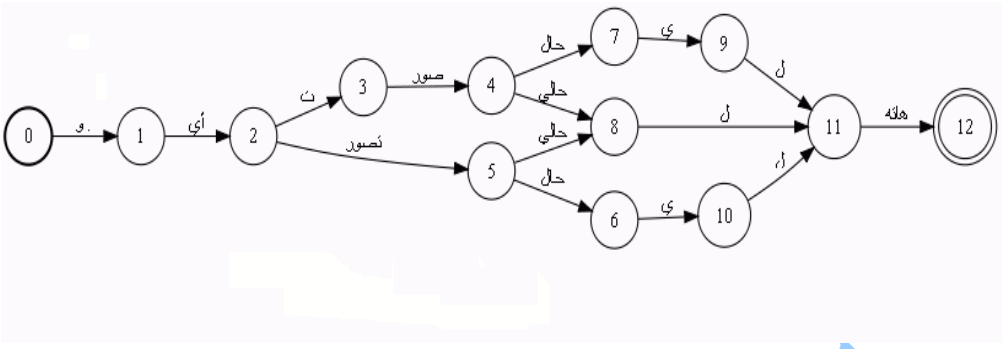

Figure 2. Exemple de segmentation d'une partie d'une phrase sous forme de treillis

La solution adoptée dans cette étude consiste à créer des treillis de mots regroupant toutes les possibilités de segmentation, et à utiliser le résultat obtenu pour réaliser un modèle statistique de langage.

Nous estimons qu'à partir d'un corpus volumineux la probabilité d'une de ces solutions sera augmentée comparée aux autres solutions. La solution sera retrouvée, davantage que les autres, sous des formes, qui permettront à l'analyseur morphologique de la déduire facilement. Le passage de mot brut au modèle prefix-stem-suffix donnera lieu à une seule solution sans aucune ambigüité.

3 Les travaux réalisés

Le corpus utilisé est constitué de 100K mots. Nous commençons par l'extraction de vocabulaire de chaque corpus, et nous décomposons individuellement chaque mot, en nous appuyant sur l'analyseur morphologique ArabMorph, pour suivre le modèle prefix-stem-suffix au cas où les préfixes et les suffixes existeraient. (Un exemple est affiché dans la figure 2). A noter que, dans l'exemple le signe '#' indique qu'un morphème est un préfixe, et le signe '+' indique qu'un morphème est un suffixe. La phrase est alors reconstruite mais chaque mot est remplacé par ledit modèle. Dans le cas où l'analyseur revoit plusieurs solutions pour un mot, toute la phrase est reconstruite et l'ensemble de solutions seront mise entre des "{}" séparés par des virgules.

Par programmation nous traitons le fichier résultat pour construire un fichier décrivant pour chaque phrase l'automate à état fini (FSM) (voir figure 3) ; nous employons ensuite l'outil lattice-tool de SRILM (Stolcke, 2002) pour créer le modèle de langage.

4 Résultats et analyse

Le corpus de test est constitué de 7K mots. Chaque mot est décomposé, pour suivre le modèle *prefix-stem-suffix*, mais cette fois, si l'analyse morphologique conduit à plusieurs solutions, la bonne, a été choisit manuellement. Puis avec l'outil ngram-count de SRILM nous créons les modèles 3-grames.

Le calcule de la perplexité n'est pas possible, nous comptons le recouvrement des N-grams présentés dans le corpus de test dans le modèle de langage réalisé pour l'apprentissage.

N-grams	% de recouvrement
Unigramme	70%
Bi-gramme	45%
3-gramme	39%

Tableau 4. Pourcentage N-gramme couverte par le modèle de langage réalisé

Le test a montré que le modèle réalisé recouvre une bonne partie du corpus de test, sachant que le corpus test ne contient que les bonnes solutions. Nous considérons cependant, que le vrai jugement sera fixé lorsque ce modèle sera introduit dans sur un système de reconnaissance automatique de la parole.

Conclusion

Dans cet article nous avons proposé une nouvelle approche pour remédier à la complexité morphologique de la langue Arabe et aux limites de ses analyseurs morphologiques. La solutions proposée, peut remédier à la complication de la segmentation des mots Arabe et peut même être appliqué à d'autres domaines. Notamment dans le domaine de la voyellation automatique de cette langue et peut aussi s'appliquer à la traduction automatique.

Nous estimons cependant, qu'il s'agit d'un travail préliminaire effectué sur un petit corpus. Nous projetons, dans la prochaine étape, d'élargir le corpus d'apprentissage et de le tester sur un système de

reconnaissance automatique de la parole. Ce travail est une étape qui constitue un pas de plus dans la réalisation de notre projet de base : en l'occurrence la réalisation d'un système de reconnaissance automatique de la parole Arabe [SC10].

Références et bibliographiques

[SC10] **Ali Sadiqui, Nouredine Chenfour -** *Réalisation d'un système de reconnaissance automatique de la parole arabe basé sur CMU Sphinx*, article publié sur « Annals. Computer Science Series » Tome 8, Avril 2010**.**

[X+06] **Bing Xiang, Kham Nguyen, Long Nguyen, Richard Schwartz, John Makhoul** - *Morphological decomposition for Arabic broadcast news transcription*. In Proc. ICASSP 2006, pages 1089–1092, 2006.

[Hei08] **Ilana Heintz** - *Arabic language modeling with finite state transducers*. In ACL 2008, Columbus, OH, 2008

[K+06] **Katrin Kirchhoff, Dimitra Vergyri, Kevin Duh, Jeff Bilmes, Andreas Stolcke** - *Morphology-based language modeling for conversational Arabic speech recognition*. Computer Speech and Language, 2006.

[MSL08] **Karima Meftouh, Kamel Smaili, Mohamed-Tayeb Laskri** – *Arabic statistical language modelling*. JADT: 9es Journées internationales d'Analyse statistique des Données Textuelles, 2008.

[RJ93] **L. Rabiner, B.H. Juang** – *Fundamental Of Speech Recognition* (pp. 450). New Jersey: Prentice Hall, 1993.

[M+09] **Mohsen Moftah, Waleed Fakhr, Sherif Abdou, Mohsen Rashwan** - *Stem-based arabic language models experiments*, Proceedings of the Second International Conference on Arabic Language Resources and Tools, avril 2009.

[LBC04] **Xiaoyong Liu, W. Bruce Croft** - *Statistical Language Modeling For Information Retrieval*. Center for Intelligent Information Retrieval (2004).