RECITAL 2005, Dourdan, 6–10 juin 2005

Etiquetage morpho-syntaxique des textes arabes par modèle de Markov caché

Abdelhamid EL JIHAD (1), Abdellah YOUSFI (2)
(1),(2) Institut d'études et de recherches pour l'arabisation
Université Mohamed V, Rabat, Maroc
(1) eljihad@ifrance.com
date de soutenance prévue : 2007
(2) yousfi240ma@yahoo.fr
date de soutenance : 19 juin 2001

Mots-clefs – Keywords

Corpus, jeu d'étiquettes, Etiquetage morpho-syntaxique, texte arabe, modèle de Markov caché
Corpus, the set of tags, the morpho-syntactic tagging, arabic text, Hidden Markov Model

Résumé - Abstract

L'étiquetage des textes est un outil très important pour le traitement automatique de langage, il est utilisé dans plusieurs applications par exemple l'analyse morphologique et syntaxique des textes, l'indexation, la recherche documentaire, la voyellation pour la langue arabe, les modèles de langage probabilistes (modèles n-classes), etc.
Dans cet article nous avons élaboré un système d'étiquetage morpho-syntaxique de la langue arabe en utilisant les modèles de Markov cachés, et ceci pour construire un corpus de référence étiqueté et représentant les principales difficultés grammaticales rencontrées en langue arabe générale.
Pour l'estimation des paramètres de ce modèle, nous avons utilisé un corpus d'apprentissage étiqueté manuellement en utilisant un jeu de 52 étiquettes de nature morpho-syntaxique. Ensuite on procède à une amélioration du système grâce à la procédure de réestimation des paramètres de ce modèle.

The tagging of texts is a very important tool for various applications of natural language processing : morphological and syntactic analysis of texts, indexation and information retrieval, vowelling of arabic texts, probabilistic language model (n-class model).
In this paper we have used the Hidden Markov Model (HMM) to tag the arabic texts. This system of tagging is used to build a large labelled arabic corpus. The experiments are carried in the set of the labelled texts and the 52 tags of morpho-syntactic nature, in order to estimate the parameters of the HMM.

1 Introduction

Le développement des corpus électroniques a bénéficié ces dernières années d'un appui vigoureux et un soutien financier important, de la communauté du traitement automatique des langues naturelles, qui voit là une étape indispensable pour la mise au point de systèmes de TAL robustes. Aujourd'hui de vaste corpus de textes électroniques étiquetés sont disponibles et sont majoritairement de langue anglaise. Ceci a permis l'essor considérable des traitements automatiques concernant cette langue; des outils d'interrogation de ces corpus ainsi que des outils d'annotations proprement dits (étiqueteurs, analyseurs syntaxique, etc.) se répandent. Leurs équivalents en français commence à apparaître également [Habert et al 1997].

Pour la langue arabe, il n'existe pas à ce jour de corpus étiqueté aisément disponible. Par conséquent les recherches linguistiques qui ont recours à des corpus étiquetés sont donc encore rares. Motivé par ce manque, l'Institut d'Etudes et de Recherches pour l'Arabisation (IERA) a entrepris un projet de recherche dont l'objectif est la constitution d'un corpus de référence étiqueté et représentant les principales difficultés grammaticales rencontrées en langue arabe générale. La disponibilité de ce corpus à l'institut, va donner le coup d'envoi aux divers travaux de recherches linguistiques qui utilisent les corpus étiquetés. Un corpus étiqueté est un corpus dans lequel on associe à des segments de textes (le plus souvent des mots) d'autres informations de quelque nature qu'elle soit morphologique, syntaxique, sémantique, prosodique, critique, etc [Veronis 2000][Vergne et al 1998].

En particulier, dans la communauté du traitement automatique des langues naturelles, quand on parle de corpus étiqueté on fait référence le plus souvent à un document où chaque mot possède une étiquette morpho-syntaxique et une seule.
L'étiquetage morpho-syntaxique automatique est un processus qui s'effectue généralement en trois étapes [Minh et al 2003][Rajman et al 2000]: la segmentation du texte en unités lexicales, l'étiquetage à priori, la désambiguïsation qui permet d'attribuer, pour chacun des unités lexicales et en fonction de son contexte, l'étiquette morpho-syntaxique pertinente.
La taille du jeu d'étiquettes, la taille du corpus d'apprentissage sont autant de facteur importants pour une bonne performance du système d'étiquetage [Chanod 1995][Claud 1995].
En général, il existe deux méthodes pour l'étiquetage morpho-syntaxique :
- Méthode à base de règles [Claud 1995][Bril 1992].
- Méthode probabiliste.
Dans cet article nous avons utilisé la deuxième approche.

2 Méthode probabiliste

Le choix de l'étiquette la plus probable en un point donné se fait au regard de l'historique des dernières étiquettes qui viennent d'être attribuées. En général cet historique se limite à une ou deux étiquettes précédentes. Cette méthode suppose qu'on dispose d'un corpus d'apprentissage qui doit être d'une taille suffisante pour permettre une estimation fiable des probabilités [Habert et al 1997].
Soit $Ph = w_1...w_T$ une phrase constituée des mots $w_1, ..., w_T$, $E = \{et_1, ..., et_N\}$ un jeu d'étiquettes.
L'étiquetage morpho-syntaxique de la phrase Ph par des étiquettes appartenant à E et s'appuyant

sur l'approche probabiliste , consiste à trouver l'ensemble d'étiquettes $et^*{}_1...et^*{}_T$ associés à la phrase Ph tel que :

$$et^*{}_1...et^*{}_T = arg \max_{et_1...et_T} Pr(w_1...w_T, et_1...et_T) \quad (1)$$

Pour faciliter la résolution de ce problème on utilise les modèles de Markov cachés d'ordre 1.

3 Etiquetage morpho-syntaxique par modèle de Markov caché d'ordre 1

Un modèle de Markov caché d'ordre 1 est un double processus $(X_t, Y_t)_{t\geq 1}$ avec :
• X_t est une chaîne de Markov d'ordre 1 à valeur dans un ensemble d'états fini $Q = \{q_1, ..., q_N\}$, X_t vérifie :
$Pr(X_{t+1} = q_j/X_1 = q_1, ..., X_t = q_i) = Pr(X_{t+1} = q_j/X_t = q_i) = a_{ij}$.
$Pr(X_1 = q_i) = \pi_i$, $i = 1, ..., N$.
a_{ij} est la probabilité de transition entre les états q_i et q_j.
π_i est la probabilité que l'états q_i est un état initial.
• Y_t est un processus observable à valeurs dans un ensemble mesurable Y, Y_t vérifie :
$Pr(Y_t = y_t/X_1 = q_1, ..., X_t = q_i, Y_1 = y_1, ..., Y_{t-1} = y_{t-1}) = Pr(Y_t = y_t/X_t = q_i) = b_i(y_t) = b_{it}$.
b_{it} est la probabilité d'émission de l'observation y_t à partir de l'état q_i.
Dans la suite on supposera que le double processus :
$X_t = et_{i_t}$ représentant les étiquettes appartenant à l'ensemble E,
$Y_t = w_t$ représentant les mots de notre vocabulaire $V = \{w_1, ..., w_L\}$,
est un modèle de Markov caché d'ordre 1.
Remarque :
Ce modèle est défini entièrement par un vecteur de paramètres noté $\lambda = (\Pi, A, B)$.
• $\Pi = \{\pi_1, ..., \pi_N\}$ l'ensemble des probabilités initiales.
• $A = (a_{ij})_{1\leq i,j\leq N}$ la matrice des probabilités de transition entre les étiquettes.
• $B = (b_{it})_{1\leq i\leq N}$ et $1 \leq t \leq L$: la matrice des probabilités d'émission des mots à partir des étiquettes.

4 Procédure d'apprentissage (Estimation des paramètres)

L'apprentissage est une opération nécessaire pour un système de reconnaissance de formes (en particulier le système d'étiquetage), il permet d'estimer les paramètres du modèle $\lambda = (\Pi, A, B)$. Un apprentissage incorrect ou insuffisant diminue la performance du système d'étiquetage. Pour préparer le corpus d'apprentissage, on procède par approximations successives. Un premier corpus d'apprentissage, relativement court, permet d'étiqueter un corpus beaucoup plus important. Celui-ci est corrigé, ce qui permet de réestimer les probabilités, il sert donc à un second apprentissage, et ainsi de suite.
En général il existe trois méthodes d'estimation de ces paramètres[1] :
• L'estimation par maximum de vraisemblance (Maximum Likelihood Estimation), elle est réalisée par l'algorithme de Baum-Welch [Baum 1972] ou l'algorithme de Viterbi [Celeux 92].

[1] Pour plus de détaille sur ces formule voir [Yousfi 2001]

• L'estimation par maximum a posteriori [John Arice].
• L'estimation par maximum d'information mutuel [Bahl et al 86,87][Kapadia 93].
Dans notre cas nous avons utilisé l'estimation par maximum de vraisemblance car c'est la plus utilisée et la plus facile à calculer.
Alors si on prend un ensemble d'apprentissage $R = \{Ph_1, ..., Ph_K\}$, constitué des phrases $Ph_1, ..., Ph_K$ étiquetées manuellement, les formules d'estimation des paramètres du modèle $\lambda = (\Pi, A, B)$ sont données par :

$$a_{ij} = \frac{\sum_{n=1}^{K} \text{le nombre de fois où la transition } et_i\ et_j \text{ est dans la phrase } Ph_n}{\sum_{n=1}^{K} \text{le nombre de fois où l'état } et_i \text{ est atteint le long de la phrase } Ph_n}$$

$$\pi_i = \frac{\sum_{n=1}^{K} \delta[\text{l'étiquette } et_i \text{ est un état initial dans la phrase } Ph_n]}{K}$$

$$b_{it} = \frac{\sum_{n=1}^{K} \text{le nombre de fois où le mot } w_t \text{ à l'étiquette } et_i \text{ le long de la phrase } Ph_n}{\sum_{n=1}^{K} \text{le nombre de fois où l'état } et_i \text{ est atteint le long de la phrase } Ph_n}$$

avec :
$\delta[x] = \begin{cases} 1 & \text{si l'événement x est vrai} \\ 0 & \text{sinon} \end{cases}$

5 Etiquetage automatique par algorithme de Viterbi

Pour un calcul plus rapide du chemin optimal[2] dans la formule (1) nous avons utilisé l'algorithme de Viterbi [For 73].
On note par :

$$\delta_t(et_j) = \max_{et_{i_1}...et_{i_t}} Pr(w_1...w_t, et_{i_1}...et_{i_t})$$

avec $et_{i_t} = et_j$.
Cette formule devient [Yousfi 2001]:

$$\delta_t(et_j) = \max_{et_i} \delta_{t-1}(et_i).a_{ij}.b_j(w_t)$$

On calcule cette formule pour toutes les valeurs $t = 1, ..., T$ et $j = 1, ..., N$.
Enfin le chemin optimal est obtenu à l'aide d'un calcul récursif sur cette formule.

6 Expérimentation

6.1 Données d'apprentissage

Le travail expérimental a été réalisé en trois grandes étapes :
1) étape de définition du jeu d'étiquettes et de construction de corpus d'apprentissage.
La définition de notre propre jeu d'étiquettes morpho-syntaxique a été particulièrement délicate, cette phase a été réalisée en collaboration avec des linguistes pour satisfaire au besoin des projets en cours de réalisation à IERA. Ce jeu d'étiquettes est constitué de 52 étiquettes de nature

[2]Nous cherchons ce chemin dans un réseau d'étiquettes. Ce réseau est construit de tel façon à ce que pour une phrase donnée, chaque chemin de ce réseau correspond à la probabilité que cette phrase à les étiquettes de ce chemin ($Pr(w_1...w_t, et_{i_1}...et_{i_t})$). Le chemin associé à la probabilité maximale est nommé chemin optimal.

morpho-syntaxique (comme par exemple ism-faail, ism-mafaoul, harf nasb,...).
Le corpus d'apprentissage est constitué d'un ensemble de phrases représentant les principales règles morphologiques et syntaxiques utilisées en langue arabe générale. Ce corpus a été étiqueté manuellement par un linguiste.
2) étape d'estimation des paramètres du modèle de Markov caché.
3) étape d'étiquetage automatique et réestimation des paramètres du modèle de Markov caché.
Pour réaliser ces deux dernières étapes, nous avons développé une application en langage C, comportant deux modules, module d'apprentissage et module d'étiquetage automatique qui permet d'étiqueter automatiquement un corpus brut, ce dernier est corrigé manuellement pour servir à une réestimation des paramètres du modèle de Markov caché.
Les programmes sont évalués sur deux versions de textes voyellé et non voyellé.

6.2 Résultats

Le taux d'erreur est mesuré sur deux ensembles :
Ensemble1 constitué des mêmes phrases que l'ensemble d'apprentissage mais sans étiquettes,
Ensemble2 constitué de phrases (sans étiquettes) différentes de celles de l'ensemble d'apprentissage.

	Ensemble1	Ensemble2
Textes voyellés	1,76%	2%
Textes non voyellés	2,5%	3%

Table 1: Les taux d'erreur d'étiquetage automatique.

On remarque que dans le cas des textes non voyellés le taux d'erreur augmente par rapport aux textes voyellés, à cause de l'augmentation de l'ambiguïté (un mot peut prendre plusieurs étiquettes). Pour le reste des erreurs, elles sont dues au manque de données d'apprentissage (il existe des mots et des transitions entre des étiquettes qui ne sont pas représentées dans le corpus d'apprentissage).

7 Conclusions et perspectives

En analysant les résultats trouvés, nous avons remarqué que la majorité d'erreurs d'étiquetage provient essentiellement du problème de manque ou d'insuffisance de données d'apprentissage.
Dans notre cas il existe deux type de problèmes de manque de données :
• un ou plusieurs mots, appartenant à la phrase à étiqueter par ce système, n'existent pas dans le lexique, c'est à dire nous n'avons pas une estimation des probabilités d'observation de ces mots dans tous les états.
• une ou plusieurs étiquettes n'ont pas de prédécesseurs dans la phrase à étiqueter automatiquement, c'est à dire nous n'avons pas une estimation des probabilités de transition de ces étiquettes vers tous les autres étiquettes du système.
Dans la suite de notre travail, nous allons proceder á deux solutions pour remedier à ces deux problèmes :

la première est d'introduire une sorte d'analyse morphologique qui s'appuit sur les formes morphologiques des mots pour pouvoir identifier les étiquettes des mots inconnus.
La deuxième est d'introduire une base de règles syntaxiques qui définie les transitions possibles entre les différents étiquettes.

Références

L.R. Bahl, P.F. Brown, P.V. de Souza & R.L. Mercer : "*Maximum mutual information estimation in hidden Markov model parameters for speech recognition* ", Proc. ICASSP, pp. 49-52, Tokyo, 1986.

L. R. Bahl, P. F. Brown, P.V De Souza and R. L. Mercer : "*Estimating HMM parameters so as to maximise speech recognition accuracy* ", Research Report RC-13121, IBM TJ Watson Research Center, 9/10/1987.

L. Baum : "*An inequality and association maximization technique in statistical estimation for probabilistic functions of Markov processes* ", Inequality, vol. 3, 1972.

G. Celux, J. Clairambault :"*Estimation de chaines de Markov cachées: méthodes et problèmes* ", Journées thématiques CNRS sur les approches markoviennes en signal et images, Septembre 1992.

Jean-Pierre Chanod and Pasi Tapanainen : "*Tagging French - comparing a statistical and a constraint-based method*", Proceeding of the seventh Conference of the European Chapter of the Association for Computatinal

Linguistics (EACL.95), Dublin, Ireland. pp.149-156, 1995.

Claude De Loupy : "*La méthode détiquetage d'Eric Brill*". Revue T.A.L, 1995, Vol.36, nř 1-2, pp.37-46

Eric Brill : "*A simple rule-based part of speech tagger*". Proceedings of the third Conference on Applied Natural Language Processing, Trento, Italy. pp.152-155. Avril 1992.

Fornay D. R. : "*The Viterbi Algorithm* ", Proc. IEEE, vol. 61, n 3, mai 1973.

Benoît Habert, Adeline Nazarenko, André Salem : "*Les linguistiques de corpus*" , Armand colin / Masson.Paris, 1997.

John Rice : "*Mathematical Statistics and data*

analysis ", page 511-540.

S. Kapadia, V. Valtchev & S.J. Young : "*MMI training for continuous phoneme recognition on the TIMIT database* ", Proc. ICASSP, pp. II.491-494, Minneapolis, 1993.

Thi Minh Huyen Nguyen, Laurent Romary, Xuan Luong Vu : "*Une étude de cas pour l'étiquetage morpho-syntaxique de textes vietnamiens*" , 5e conférence sur le traitement Automatique du Langage Naturel (TALN2003), Batz-sur-Mer, 11-14 juin, 2003.

Patrick Paroubek et Martin Rajman : "*Etiquetage morpho-syntaxique.*" , Ingénierie des langues. pp.131-150, Paris, HERMES Sciences Europe.

Jacques Vergne, Emmanuel Giguet: "*Regards théoriques sur le "Tagging"* " , 5e conférence sur le traitement Automatique du Langage Naturel (TALN98), Paris, France, 10-12 juin, 1998.

Jean Veronis : "*Annotation automatique de corpus : panorama et état de la technique*" , Ingénierie des langues. pp.111-128. Paris, HERMES Sciences Europe.

A. Yousfi : "*Introduction de la Vitesse d'élocution dans un modèle de reconnaissance automatique de la parole* " , Thèse de doctorat, 19 juin 2001.