Abdellah YOUSFI, Abdelhamid EL JIHAD,
Institut d'Etudes et de Recherches pour l'Arabisation
Université Mohamed V Soussi, Rabat, Maroc
yousfi240ma@yahoo.fr
eljihad@ifrance.com

Traitement des mots inconnus pour les systèmes d'étiquetage morpho-syntaxiques des textes arabes basés sur le modèle de Markov caché

Résumé :

L'étiquetage morpho-syntaxique des textes est un outil très important pour le traitement automatique du langage, il est utilisé dans plusieurs applications dont l'analyse morphologique et syntaxique des textes, l'indexation, la recherche documentaire, la voyellation pour la langue arabe, les modèles de langage probabiliste (modèles n-classes), etc.

Ces systèmes se basent toujours sur des lexiques de taille limitée, par conséquent ils sont incapable de traiter les mots inconnus.

Pour remédier à ce problème, nous avons élaboré dans ce papier un système d'étiquetage morpho-syntaxique se basant sur les modèles de Markov caché et intégrant les formes des mots inconnus.

Pour l'estimation des paramètres de ce nouveau modèle, nous avons utilisé un corpus d'apprentissage étiqueté manuellement et un lexique des formes des mots arabes, 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.

Introduction

L'étiquetage automatique des textes est un processus qui consiste à associer à des segments de textes (le plus souvent des mots) d'autres informations de quelque nature qu'elle soit morpho- logique, syntaxique, sémantique, prosodique, critique, etc [Veronis 2000][Vergne et al 1998]. 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. En général, il existe deux approches principales pour réaliser la désambiguïsation : les méthodes à base de règles et les méthodes probabilistes. Parmi les problèmes qui se posent dans les systèmes d'étiquetage, est celui des mots inconnus (les mots n'appartenant pas au vocabulaire du système). Tous les vocabulaires des systèmes d'étiquetage sont de taille limitée, par conséquent il y a toujours des mots que ces systèmes sont incapables de traiter.

Dans ce papier, nous avons élaboré une approche pour résoudre le problème des mots inconnus, en utilisant la notion des formes des mots. Cette approche est introduite dans le système d'étiquetage morpho-syntaxique, à base du modèle de Markov caché, développé au sein de l'IERA [El Jihad, Yousfi 2005].

L'étiquetage par 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 qui précédent. Cette méthode suppose qu'on dispose d'un corpus d'apprentis- sage d'une taille suffisante pour permettre une estimation fiable des probabilités [Habert et al 1997].

Soit P h = w1...wp une phrase constituée des mots w1, ..., wp appartenant au vocabulaire V du système, E = {et1 , ..., etN } un jeu d'étiquettes.

L'étiquetage morpho-syntaxique de la phrase P h par des étiquettes appartenant à E et s'appuyant sur l'approche probabiliste, consiste à trouver l'ensemble des étiquettes et∗ 1...et∗ p associées à la phrase P h tel que

$$et*_1 ...et*_p = \arg \max_{et1 ...etp} P r(w1 , ..., wp, et1, ..., etp) \quad (1)$$

Le problème qui se pose dans cette formule est celui des mots n'appartenant pas à V . Pour résoudre l'équation (1) en prenant en compte ce problème, nous avons adapté le modèle de Markov caché en introduisant la notion des formes de ces mots inconnus.

Etiquetage morpho-syntaxique par modèle de Markov caché d'ordre 1 avec utilisation des formes

Un modèle de Markov caché d'ordre 1 en prenant en compte les formes des mots, est un processus (Xt, Yt, Zt)t≥1 associé à l'ensemble des paramètres λ = (Π, A, B, D) :

- • Π = {π1, ..., πN } l'ensemble des probabilités initiales.
- • A = (aij)1≤i,j≤N la matrice des probabilités de transition entre les étiquettes.
- • B = (bit)1≤i≤N et 1 ≤ t ≤ L : la matrice des probabilités d'émission des mots à partir des étiquettes.
- • D = (dit)1≤i≤N et 1 ≤ t ≤ L : la matrice des probabilités d'émission des formes à partir des étiquettes.

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 λ =(Π, A, B, D). 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.

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

Pour un calcul plus rapide du chemin optimal, nous avons adapté l'algorithme de Viterbi [For 73] pour résoudre l'équation (1).

Expérimentation

Le travail expérimental a été réalisé en quatre grandes étapes :

- • Etape de définition du jeu d'étiquettes et de construction du corpus d'apprentissage. La définition de notre propre jeu d'étiquettes morpho-syntaxiques 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 morpho- syntaxique.
- 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 éti- queté manuellement par un linguiste.
- • Etape de construction de la base de données des formes des mots. Cette base est constituée de 3800 formes non voyellées, et elle est générée à partir d'un générateur morphologique des verbes développé au sein de l'IERA. Cette base est ensuite enrichie par les formes des noms dérivés.
- • Etape d'estimation des paramètres du modèle de Markov caché adapté.
- • Etape d'étiquetage automatique et réestimation des paramètres du modèle de Markov caché. Le taux d'erreurs est mesuré sur un ensemble de test contenant 500 phrases non voyellées.

	Ensemble test
Ancien système (MMC)	4%
Nouveau système (MMC adapté)	2,6%

Table 1 : Les taux d'erreurs sur l'ensemble de test pour l'ancien et le nouveau système. Nous constatons que notre modèle a apporté une amélioration de 1.4% du taux d'erreurs pour cet ensemble test. Ceci est dû au fait que ce nouveau système a réussi à étiqueter correctement les phrases contenant des mots inconnus.

Références

1. Fornay D. R. : "The Viterbi Algorithm ", Proc. IEEE, vol. 61, n 3, mai 1973.
2. Benoît Habert, Adeline Nazarenko, André Salem : "Les linguistiques de corpus" , Armand colin / Masson.Paris, 1997.
3. Thi Minh Huyen Nguyen, Laurent Romary, Xuan Luong Vu : "Une étude de cas pour l'étiquetage morpho-syntaxique de textes vietnamiens" , 5◦ conférence sur le traitement Automatique du Langage Naturel (TALN2003), Batz-sur-Mer, 11-14 juin, 2003.
4. Jacques Vergne, Emmanuel Giguet : "Regards théoriques sur le "Tagging" " , 5◦ 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.
5. Yousfi : "Introduction de la Vitesse d'élocution dans un modèle de reconnaissance automatique de la parole "Thèse de doctorat, 19 juin 2001.4