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

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

1. 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 morphologique, syntaxique, sémantique, prosodique, critique, etc [1][2].

L'étiquetage morpho-syntaxique automatique est un processus qui s'effectue généralement en trois étapes [3][4] :

1) la segmentation du texte en unités lexicales.

2) L'étiquetage à priori qui consiste à attribuer, pour chacune des unités lexicales identifiées, l'ensemble des étiquettes morpho-syntaxiques possibles. Cette opération qui est produite par un programme qu'on appelle un étiqueteur (tagger), peut se faire par consultation d'un lexique où chaque forme est suivie d'une liste de catégories soit par analyse morphologique soit par combinaison des deux.

3) la désambiguïsation qui permet d'attribuer, pour chacune des unités lexicales et en fonction de son contexte, l'étiquette morpho-syntaxique pertinente.

En général, il existe deux approches principales pour réaliser cette tâche : 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 [5].

2. Les formes des mots arabes

La langue arabe a une caractéristique particulière, les verbes et les noms dérivés peuvent être regroupés selon ce qu'on appelle les formes. La construction d'une forme d'un mot se fait selon la procédure suivante [6]: on extrait du mot les lettres radicales composant sa racine, ensuite on les remplace dans ce mot selon la méthode suivante : la première lettre radicale est remplacée par "ف", la deuxième par "ع", la troisième par "ل" et la quatrième par "ل".

Exemple :

La forme du mot "اِنتقل" est "اِفتعل".

La forme du mot "تدحرجَا " est "تفعلّلا".

Pour la langue arabe, chaque verbe voyellé possède une forme. Pour réaliser la phase de recherche de la forme, on a développé une distance D qui permet de mesurer le degré de similarité exacte entre le verbe à traiter et les formes de la base. Ensuite, on choisi la forme ayant la distance maximale. Si on note par $F=\{f_1,f_2 ,..., f_N\}$ l'ensemble de toutes les formes de la base et v le verbe ayant la forme *f*, cette dernière est donnée par la formule suivante :

$$f = \arg\max_{i=1,...,N} D(v, f_i)$$

3. 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'apprentissage d'une taille suffisante pour permettre une estimation fiable des probabilités [7].

Soit $Ph = w_1...w_p$ une phrase constituée des mots $w_1,...,w_p$, appartenant au vocabulaire V du système $E = \{et_1,...,et_N\}$ 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 $et_1^*,...,et_p^*$ associés à la phrase *Ph* tel que:

$$et_1^*,...,et_p^* = \arg\max_{et_1,...,et_p} \Pr(w_1,...,w_p,et_1,...,et_p) \quad (1)$$

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.

4. 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 $(X_t,Y_t, Z_t)_{t\geq 1}$ 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:

$$\Pr(X_{t+1} = q_j / X_1 = q_1,..., X_t = q_i) =$$
$$= \Pr(X_{t+1} = q_j / X_t = q_i) = a_{ij}$$

$$\Pr(X_1 = q_i) = \pi_i \quad i = 1,...,N$$

- a_{ij} est la probabilité de transition entre les états q_i et q_j
- π_i est la probabilité que l'état 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.

- Z_t est un processus observable à valeurs dans un ensemble mesurable Z, Z_t vérifie:

$$\Pr(Z_t = z_t / X_1 = q_1,..., X_t = q_i, Z_1 = z_1,..., Z_{t-1} = z_{t-1}) =$$
$$= \Pr(Z_t = z_t / X_t = q_i) = d_i(z_t) = d_{it}$$

- d_{it} est la probabilité d'émission de l'observation z_t à partir de l'état q_i

Dans la suite on supposera que le processus:

$X_t = et_{it}$ représentant les étiquettes appartenant à l'ensemble E ,

$Y_t = w_t$ représentant les mots du vocabulaire $V = \{w_1,...,w_L\}$,

$Z_t = f_t$ représentant les formes des mots du vocabulaire V,

est un modèle de Markov caché d'ordre 1.

Remarque:

Ce modèle est défini entièrement par un vecteur de paramètres noté $\lambda = (\Pi,A,B,D)$.

- $\Pi = \{\pi_1,...,\pi_N\}$ l'ensemble des probabilités initiales.
- $A = (a_{ij})_{1\leq i,j\leq N}$: la matrice des probabilités de transition entre les étiquettes.
- $B = (b_{it})_{1\leq i\leq N \text{ et } 1\leq t\leq L}$: la matrice des probabilités d'émission des mots à partir des étiquettes.
- $D = (d_{it})_{1\leq i\leq N \text{ et } 1\leq t\leq L}$: la matrice des probabilités d'émission des formes à partir des étiquettes.

5. 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 $\lambda = (\Pi,A,B,D)$. 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.

En général il existe trois méthodes d'estimation de ces paramétrer [8].

- L'estimation par maximum de vraisemblance (Maximum Likelihood Estimation). Elle est réalisée par l'algorithme de Baum-Welch [9] ou l'algorithme de Viterbi [10].
- L'estimation par maximum a posteriori [11].
- L'estimation par maximum d'information mutuel [12].

Dans notre cas nous avons utilisé l'estimation par maximum de vraisemblance.

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

$\lambda = (\Pi,A,B,D)$ par cette méthode, sont données par:

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

$$\pi_i = \frac{\sum_{n=1}^{k} \delta[et_i \text{ état initial de ph}_n]}{k}$$

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

$$d_{it} \frac{\sum_{n=1}^{k} \text{le nombre de fois où la forme } f_t \text{ a l'étiquette } et_i \text{ le long de la phrase } ph_n}{\sum_{n=1}^{k} \text{le nombre de fois où l'état } et_i \text{ est atteint le long de la phrase } ph_n}$$

avec:

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

6. Etiquetage automatique par algorithme de viterbi adapté

Pour un calcul plus rapide du chemin optimal, nous avons adapté l'algorithme de Viterbi [13].pour résoudre l'équation (1).

On note par :

$$\delta_t(et_j) = \max_{et_{i_1} \ldots et_{i_t}} \Pr(w_1 \ldots w_t, et_{i_1} \ldots et_{i_t}) \quad (2)$$

$$avec \quad et_{i_t} = et_j$$

Pour résoudre le problème des mots inconnus, nous avons introduit le processus des formes de ces mots dans la formule (2). Cette formule devient [8].:

$$\delta_t(et_j) = \begin{cases} \max_{et_i} \delta_{t-1}(et_i).a_{ij}.b_j(w_t) & si \quad w_t \in V \\ \max_{et_i} \delta_{t-1}(et_i).a_{ij}.d_j(f_t) & \text{sinon} \end{cases}$$

On calcule cette formule pour toutes les valeurs $t = 1,...,T$ et $j=1,...,N$.

En fin le chemin optimal est obtenu à l'aide d'un calcul récursif sur cette formule.

7. Experimentation

Données d'apprentissage

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-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 morpho-syntaxique .

signification	étiquettes
[illegible]	[illegible]
[illegible]	[illegible]
[illegible]	[illegible]
[illegible]	[illegible]
[illegible]	[illegible]
[illegible]	[illegible]
[illegible]	[illegible]
[illegible]	[illegible]

Exemple de quelques étiquettes morpho-syntaxique utilisé dans notre système

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.

./. □.//□□□□□□ □.□/□ ./.□/□□□□

./. □.//□□□□□□ □.□/□ □.□/□□ ./.□/□□□□

./. □.//□□□□ ./□/□ □.//□□□□□□ □.□/□ ./.□/□□

./. ./// ./.□/□□□ □.□/□□□□

Exemple d'un extrait de notre corpus d'apprentissage

- Etape de construction de la base de données des formes des mots. Cette base est constitué 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 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.

Pour réaliser ces deux dernières étapes, nous avons développé une application en langage C, comportant trois modules :

* module de détermination de la forme d'un mot donné,
* module qui permet de réaliser la phase d'apprentissage,
* module d'étiquetage automatique d'un corpus brut. Ce dernier est corrigé manuellement pour servir à une ré estimation des paramètres du modèle de Markov caché adapté.

Les programmes sont évalués sur la base d'une version de texte non voyellé.

Résultats

Le taux d'erreur est mesuré sur un ensemble de test contenant 500 phrases non voyellées.

	Ensemble test
Ancien système(MMC)	4%
Nouveau système(MMC adapté)	2.6%

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.

8 Conclusions et perspectives

En analysant les résultats dégagés, nous avons remarqué que les erreurs d'étiquetage proviennent essentiellement du fait qu'une ou plusieurs étiquettes n'ont pas de prédécesseurs dans la phrase à étiqueter automatiquement, c'est à dire que nous n'avons pas une estimation des probabilités de transition de ces étiquettes vers toutes les autres. De même, le problème de l'incapacité de notre système à trouver toutes les formes possibles d'un mot inconnu (dans le cas où le mot admet plusieurs formes ("نجار" → "فعال", "نفعل")) reste à résoudre.

Comme perspective de notre travail, nous comptons améliorer l'algorithme de recheche des formes des mots afin de trouver toutes les formes possibles d'un mot donné. Nous comptons introduire également une sorte d'analyse syntaxique dans notre système pour remedier au problème des transitions.

Références :

1 Veronis, J. 2000. «Annotation automatique de corpus: panorama et état de la technique», Ingénierie des langues. Paris, HERMES Sciences Europe. pp.111-128.

2 Vergne,J. et Emmanue.l G.. «Regards théoriques sur le Tagging », », (TALN1998), Paris, France.

3 Thi Minh, Hu., Laurent, R. et Xuan L. 2003. «Une étude de cas pour l'étiquetage morpho-syntaxique de textes vietnamiens », (TALN2003), Batz-sur Mer.

4 Paroubek, P. et Martin, R. «Etiquetage morpho-syntaxique. », Ingénierie des langues. Paris, HERMES Sciences Europe. pp.131-150.

5 EL JIHAD A.et Yousfi A. "Etiquetage morpho-syntaxique des textes arabes par modèle de Markov caché", (RECITAL 2005)., 06-10 Juin 2005 Dourdan, France.

6 Al Ghalayni, M. 2000. جامع الدروس العربية. المكتبة العصرية

7 Habert, B., A, Nazarenko, et A, Salem. 1997. Les linguistiques de corpus, Armand colin / Masson.Paris.

8 Yousfi, A. 2001. «Introduction de la Vitesse d'élocution dans un modèle de reconnaissance automatique de la parole », Thèse de doctorat, université : Mohamed premier, Oujda, Maroc. 115p.

9 Baum,L. 1972. «An inequality and association maximization technique in statistical estimation for probabilistic functions of Markov processes » Inequality, vol. 3.

10 Celux, G.et J. Clairambault. 1992. «Estimation de chaînes de Markov cachées: méthodes et problèmes », Journées mathématiques CNRS sur les approches markoviennes en signal et images.

11 John. R. Mathematical Statistics and data analysis. pp 511-540.

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

13 Fornay,D. R. 1973. «The Viterbi Algorithm », Proc. IEEE, vol. 61, n° 3