Examen Master Informatique – Intelligence Artificielle et Décision
21 – 11 – 2007
Durée 2 h, Documents autorisés

Traduction automatique

On s'intéresse à la traduction automatique par des méthodes statistiques et plus précisément au problème de l'alignement de phrases. On suppose que l'on dispose de corpus alignés, constitués de phrases, en 2 langues, et l'on veut apprendre à traduire ces phrases d'une langue à l'autre. Pour cela, on va chercher la correspondance entre les deux phrases.

On note *E* une phrase anglaise, *F* une phrase française, *P(E|F)* est la probabilité que *E* soit la traduction anglaise de la phrase française *F* (même chose pour *P(F|E)*).
On suppose que les mots de *F* ont *au plus* un mot associé dans *E*. Certains mots de *F* n'ont donc aucun mot associé dans *E* et plusieurs mots de *F* peuvent avoir le même associé dans *E*.
On note $F = f_1^m = f_1 ... f_m$, $E = e_0^n = e_0 e_1 ... e_n$, respectivement une phrase *F* de *m* mots, et une phrase *E* de *n* mots avec e_0 le symbole « mot vide » auquel pourront être associés des mots de *F*,
On définit un alignement $A = a_1^m = a_1 ... a_m$, entre *E* et *F* de la façon suivante : $a_j \in \{0,1,\ldots,n\}$, $a_j = i$ signifie que le jème mot de *F* est connecté au ième mot de *E*, $a_j = 0$ si il n'est connecté à aucun mot.
Exemples :
F = « jean aime Marie », E = John loves Mary », A = (1 2 3)
F = « Le chien est battu par Jean », E = « John does beat the dog », A = (4 5 3 3 0 1)

Questions

1. On s'intéresse aux probabilités de traduction *P(F| E)*.
1.1 Montrer

$$P(F, A \mid E) = \prod_{j=1}^{m} P(a_j \mid a_1^{j-1}, f_1^{j-1}, E) P(f_j \mid a_1^{j}, f_1^{j-1}, E)$$

1.2 Interpréter l'équation précédente en termes de modèle génératif du processus de traduction.

2. On fait maintenant les hypothèses suivantes

$$P(a_j \mid a_1^{j-1}, f_1^{j-1}, E) = \frac{1}{n+1}$$
$$P(f_j \mid a_1^{j}, f_1^{j-1}, E) = P(f_j \mid e_{a_j})$$

2.1 Interpréter ces hypothèses
2.2 Exprimer *P(F, A| E)* en prenant en compte ces hypothèses.
2.3 Montrer

$$P(F \mid E) = K \sum_{a_1=0}^{n} \ldots \sum_{a_m=0}^{n} \prod_{j=1}^{m} P(f_j \mid e_{a_j})$$

3. Estimation des paramètres
Le modèle que l'on vient d'introduire a comme *paramètres* les probabilités de traduction *P(f | e)* pour l'ensemble des mots *f* et *e* du français et de l'anglais.
Pour apprendre un modèle de traduction, on va chercher à estimer ces paramètres sur un corpus de phrases alignées Français, Anglais.
On considère, pour simplifier les notations, un seul couple de phrases (*E,F*).
Pour l'estimation des probabilités de traduction $P(f_j \mid e_{a_j})$, on va chercher à résoudre le problème suivant :

$$\begin{cases} \text{Maximiser } P(F|E) \\ \text{Sous les contraintes } \sum_f P(f \mid e) = 1, \quad \forall e \end{cases} \qquad (1)$$

La somme porte sur tous les mots *f* possibles et on a autant de contraintes que de mots *e*.

3.1 A quel type de problème rencontré dans le cours cela vous fait il penser ?
3.2 Ce problème est il soluble directement analytiquement ? Expliquer.

On propose de résoudre le problème en utilisant un algorithme itératif EM.
On définit la fonction auxiliaire $Q(\theta, \hat{\vartheta}) = \sum_{A \in Alignements(F,E)} \log(P_\theta(F, A \mid E)) P_{\hat{\vartheta}}(A \mid F, E)$
« *Alignements (F,E)* » désigne l'ensemble de tous les alignements possibles de F et *E*.
Dans cette expression, $\hat{\theta}$ désigne les paramètres courants (valeurs initiales au début de l'algorithme ou valeurs estimées à l'étape courante), θ désigne les nouveaux paramètres que l'on veut estimer. P_θ désigne une probabilité fonction de θ et $P_{\hat{\theta}}$ est l'estimateur courant.
3.3 Que représente cette fonction ? Expliquer ce que sont les deux quantités sous la somme.
3.4 Donner l'expression de *Q* avec les hypothèses faites précédemment.
On admet le résultat suivant :

$$Q(\theta, \hat{\vartheta}) = \sum_{j=1}^{m} \sum_{k=1}^{n} P_{\hat{\theta}}(k \mid j, F, E) \log P_{\vartheta}(f_j \mid e_k) + K$$

Dans cette expression, $P_{\hat{\theta}}(k \mid j, F, E)$ désigne la probabilité, calculée avec les paramètres courants, que le kième mot de *E*, e_k, soit aligné avec le jème de *F*, f_j, *K* sera considéré comme une constante.
On va alors remplacer le problème (1) par le problème suivant :

$$\begin{cases} \text{Maximiser } Q \\ \text{Sous les contraintes } \sum_f P(f|e) = 1, \quad \forall e \end{cases} \qquad (2)$$

3.5 Donner le Lagrangien pour ce problème.

3.6 Montrer que la dérivée du Lagrangien par rapport à la variable *P(f|e)* s'écrit :

$$\sum_{j=1}^{m} \sum_{k=0}^{n} P_{\hat{\theta}}(k|j,F,E)\delta(f,f_j)\delta(e,e_k)\frac{1}{P_\theta(f|e)} - \lambda_e$$

où $\delta(a,b) = 1$ si $a = b$ et 0 sinon

3.7 On admet $P_{\hat{\vartheta}}(k|j,F,E) = \dfrac{P_\vartheta(f_j|e_k)}{\sum_{i=0}^{n} P_\theta(f_j|e_i)}$, en déduire une expression pour $P_\vartheta(f\,|e\,)$

3.8 Proposer un algorithme pour résoudre le problème (2).

5. Article associé : « Discriminative Word Alignment with Conditional Random Fields »

5.1 Quelle est la différence essentielle entre la méthode proposée dans ce papier et celle étudiée précédemment ? Dans quelles classes de modèles pouvez-vous positionner ces deux approches ?

5.2 Quels sont les principaux avantages avancés pour l'approche proposée dans le papier ?

5.3 Expliquer :

- « The CRF is conditionned on both source and target sentences »
- "The model allows regularization using a prior over the parameters"

5.4 Dans la formule (1), que représentent les quantités h_k, λ_k, Z ?

5.5 Décrire le processus d'apprentissage : quelles données, quel critère, quel algorithme, etc ?

Rappel : Multiplicateurs de Lagrange

Soit un vecteur réel u, *A* et *B* deux fonctionnelles de *u* à valeurs dans *R*. On considère le problème d'optimisation suivant :

minimiser A(u) (2)
sous la contrainte B(u) = b

On note $L(u,\lambda) = A(u) + \lambda(B(u) - b)$ *le Lagrangien* associé à ce problème de minimisation.

Résultat :

Une condition nécessaire pour que *u* soit solution du problème de minimisation (2) est :

$grad_u\,(L(u,\lambda)) = grad_u(A(u) + \lambda(B(u)) = 0.$ (on considérera ici qu'elle est également suffisante).

En calculant la dérivée du Lagrangien on obtient une série d'équations avec le terme λ (une équation par composante de u). On éliminera λ en utilisant la contrainte B(u) = b.