IFT6010 Traitement Automatique des Langues Naturelles
FINAL — **Automne 2008** Philippe Langlais
Durée: 2h

Directives pédagogiques

√ Inscrivez tout de suite votre nom sur le carnet de réponses
√ Documentation autorisée, calculatrice autorisée
√ Répondre sur le carnet de réponse en indiquant toujours clairement à quelle question vous répondez
√ On vous demande de répondre de manière claire et concise
√ Le barème est donné à titre indicatif. Cet examen compte pour 25 points (26 points sont à prendre) dans la note finale à ce cours.

Réglement sur le plagiat
(extrait du règlement disciplinaire sur le plagiat ou la fraude de l'Université de Montréal)

Constitue un plagiat:

1. faire exécuter son travail par un autre
2. utiliser, sans le mentionner, le travail d'autrui
3. échanger des informations lors d'un examen
4. falsifier des documents

Le plagiat est passible de sanctions allant jusqu'à l'exclusion du programme.

Modèle de langue et lissage (5 pts)

Vos réponses dans cette question doivent être suffisamment précises pour qu'un non spécialiste du domaine puisse implémenter vos propositions.

1. Énoncez en une phrase claire la différence principale entre un modèle interpolé et un modèle de repli (backoff).

2. Dans la suite, considérez un corpus $\mathcal{T} = \{t_1, \ldots, t_N\}$ de N mots. Vous décidez de créer un modèle bigramme en combinant deux modèles selon:

$$p_{comb}(w_i|w_j) = \lambda_1 p_{MLE}(w_i|w_j) + \lambda_2 p_K(w_i)$$

   où $p_{MLE}(\bullet)$ est un modèle bigramme entraîné par maximum de vraisemblance sur le corpus $\mathcal{T}$ et $p_K(\bullet)$ est un modèle unigramme entrainé par maximum de vraisemblance sur $\mathcal{T}_\mathcal{K}$ obtenu à partir de $\mathcal{T}$ en transformant chaque mot t_i en son préfixe d'au plus K caractères. Par exemple, si $\mathcal{T} = \{le, lissage, par, troncation, est, discutable\}$ alors $\mathcal{T}_4 = \{le, liss, par, tron, est, disc\}$.

   (a) Indiquez comment calculer les deux modèles $p_{MLE}(\bullet)$ et $p_K(\bullet)$. Vous donnerez une formule pour chaque modèle que vous expliquerez brièvement.

   (b) Exprimez λ_2 en fonction de λ_1 et de comptes que vous préciserez. Vous prendrez soin d'expliquer votre démarche.

   (c) Existe-t-il des situations où p_{comb} donnera une probabilité nulle ? Proposez le cas échéant une solution y remédiant.

Grammaires (5 pts)

1. Quel est le langage reconnu par la grammaire suivante: $G_1 =< S, \{S, T\}, \{a\}, \mathcal{R} >$ où $\mathcal{R}$ est: $\{S \to Sa, S \to SaS, S \to T, T \to S, T \to \epsilon\}$. Une réponse imprécise est une réponse fausse.

2. G_1 est-elle ambiguë ? Justifiez.

3. Écrire une grammaire G_2 qui reconnaît les séquences d'au moins un chiffres sur l'alphabet $\{1, 2, 3, 4\}$ telles qu'il n'existe aucun chiffre dans la chaîne qui est précédé (directement ou pas) d'un chiffre strictement supérieur. `124`, `1111` ou `4` sont des chaînes de ce langage, alors que `2312` n'en est pas.

4. Selon la classification chomskienne, de quel type sont les grammaire G_1 et G_2 ? Quel est le type des langages associés ?

Séminaires (3 pts)

1. Résumez brièvement trois séminaires auxquels vous avez assisté dans le cadre des séminaires RALI-OLST qui se tenaient les mercredis à 11h30.

Traduction (3 pts)

Considérez le modèle de transfert suivant (on suppose qu'il n'existe pas d'événement "e_0"):

	A	B	C
a	0.2	0.3	0.5
b	0.3	0.6	0.1
c	0.4	0.4	0.2

1. Dessinez l'alignement le plus probable obtenu par un modèle IBM 1 qui utilise ces probabilités de transfert pour la paire de phrases: $<A\ C,\ a\ b\ c>$

2. Quelle est la probabilité de cet alignement étant donnée la paire de phrases $<A\ C,\ a\ b\ c>$? Vous n'avez pas besoin de faire le calcul et pouvez vous contenter de l'exprimer (ex: $4 + 3$ au lieu de 7). S'il vous manque des probabilités qui ne sont pas exprimables à l'aide des probabilités ici présentes, identifiez-les et utilisez-les ensuite dans l'expression du calcul.

HMM (5 pts)

1. Vous disposez d'un bitexte de N paires de phrases $\{(e, f)\}_{i\in[1,N]}$ et souhaitez modéliser $p(e|f)$ à l'aide d'un modèle markovien d'ordre 1, afin de réaliser un système de traduction de la langue $\mathcal{F}$ vers la langue $\mathcal{E}$. Précisez les points suivants:

   (a) Quels sont les états cachés de votre modèle ?

   (b) Quelle topologie proposez-vous de donner à votre modèle ? Vous expliquerez comment la créer.

   (c) Précisez les distributions d'émission et de transition avec lesquelles décomposer $p(e|f)$.

   (d) Nommez un algorithme vu en cours permettant d'obtenir la traduction optimale selon votre modèle (supposé entraîné). Il ne vous est pas demandé de décrire l'algorithme, mais simplement d'expliquer (à l'aide d'une formule) le problème de maximisation auquel vous faites face.

   (e) Nous avons vu comment entraîner les paramètres d'un modèle de Markov par l'algorithme EM et avons discuté la sensibilité de cet algorithme aux valeurs initiales. Proposez une méthode d'initialisation des paramètres qui utilise des modèles vus en cours.

PCFG (5 pts)

Considérez la grammaire probabiliste $G =< S, \{NP, VP, NP, ART, ADJ, VB, NC, PP\}, \{la, le, belle, ferme, voile\}, \mathcal{R} >$ où $\mathcal{R}$ est:

S → NP VP `[1.0]`	ART → *la* `[0.5]`	NC → *belle* `[0.1]`
VP → PP VB `[0.6]`	ART → *le* `[0.5]`	NC → *ferme* `[0.6]`
VP → VB NP `[0.4]`	ADJ → *belle* `[1.0]`	NC → *voile* `[0.3]`
NP → ART ADJ NC `[0.3]`	VB → *ferme* `[0.2]`	PP → *le* `[0.5]`
NP → ART NC `[0.7]`	VB → *voile* `[0.8]`	PP → *la* `[0.5]`

1. Donnez un arbre d'analyse de la chaîne *la belle ferme le voile* par la grammaire G et donnez la probabilité associée (on vous demande d'exprimer la probabilité à l'aide des probabilités élémentaires, pas de faire le calcul).

2. Quelle est la probabilité de cette même chaîne selon G ? Là encore, on vous demande seulement l'expression du calcul (vous n'avez pas besoin de calculatrice).

3. Représentez la table d'analyse CYK crée lors de l'analyse de cette chaîne.