IFT6010 Traitement Automatique des Langues Naturelles
FINAL — **Automne 2006** 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

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. Vous disposez d'une collection de C documents $\{d_1, \ldots, d_C\}$ et vous souhaitez réaliser un système de recherche de documents à partir d'une requête de n mots: $r_1 \ldots r_n$. Indiquez comment vous pourriez réaliser votre système à l'aide de modèles de langue.

2. Vous souhaitez faire usage d'un modèle de langue interpolé pour modéliser une langue naturelle. L'application en faisant usage mettra votre modèle dans la situation où des événements non vus à l'entraînement devront être notés. Une technique simple consiste à ajouter dans votre vocabulaire V un mot spécial `UNK` auquel tous les mots inconnus rencontrés en test seront associés.

   (a) Indiquez une manière simple de traiter cet événement UNK dans un modèle de langue de type interpolé.

   (b) Proposez un autre mécanisme de traitement des mots inconnus qui ne stocke pas de paramètre $p(\text{UNK}|h)$ (où h désigne un historique). Votre solution doit préserver le fait que votre modèle définit bien une distribution probabiliste.

Grammaires (5 pts)

Considérez la grammaire suivante:

Ph → Sujet Verbe Comp	GNominal → Art Adj Nom	Verbe → *is* | *can* | *drink*
Ph → Verbe Comp	GNominal → Art Nom	Nom → *can* | *bottle*
Sujet → GNominal	Comp → GNominal	Pronom → *it* | *I*
Sujet → Pronom	Adj → *empty* | *full*	Art → *the*

1. Cette grammaire est-elle LL1 ? Justifiez.

2. Construisez la table d'analyse de l'algorithme d'Earley pour l'analyse de la phrase: *I drink the can*. Chaque item considéré par l'algorithme doit être indiqué, ainsi que l'opération ayant mené à sa considération.

3. Calculez `FIRST` et `FOLLOW` du non-terminal `GNominal`.

Traduction (5 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 de viterbi 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 votre calcul.

HMM et PCFG (10 pts)

Considérez le modèle markovien H dont les matrices de transition A, d'emission B et de transition initiale π sont données par:

$$A = \left[\begin{array}{c|ccccc} & s_1 & s_2 & s_3 & s_4 & s_5 \\ \hline s_1 & 0.2 & 0.5 & 0.0 & 0.0 & 0.3 \\ s_2 & 0.0 & 0.4 & 0.6 & 0.0 & 0.0 \\ s_3 & 0.0 & 0.0 & 0.0 & 1.0 & 0.0 \\ s_4 & 0.0 & 1.0 & 0.0 & 0.0 & 0.0 \\ s_5 & 0.0 & 0.8 & 0.0 & 0.0 & 0.2 \end{array}\right], B = \left[\begin{array}{c|cc} & a & b \\ \hline s_1 & 1.0 & 0.0 \\ s_2 & 1.0 & 0.0 \\ s_3 & 0.0 & 1.0 \\ s_4 & 1.0 & 0.0 \\ s_5 & 0.0 & 1.0 \end{array}\right], \pi = [0.2, 0.3, 0.1, 0.1, 0.3]$$

et la grammaire probabiliste $G =< A, \{A, B, C, D, E\}, \{a, b\}, r >$ où r est:

A $\rightarrow a$ B `[0.3]`	B $\rightarrow a$ B `[0.1]`	E $\rightarrow a$ B `[0.5]`	C $\rightarrow a$ D `[0.6]`	D $\rightarrow a$ B `[0.3]`
A $\rightarrow b$ E `[0.3]`	B $\rightarrow b$ C `[0.6]`	E $\rightarrow b$ E `[0.2]`	C $\rightarrow \epsilon$ `[0.4]`	D $\rightarrow \epsilon$ `[0.7]`
A $\rightarrow a$ A `[0.4]`	B $\rightarrow \epsilon$ `[0.3]`	E $\rightarrow \epsilon$ `[0.3]`		

1. Représentez graphiquement H.
2. Quel est le langage reconnu par H si on admet que seul s_1 n'est pas un état terminal du modèle ?
3. Quel est le langage reconnu par G ?
4. Donnez un arbre d'analyse de la chaîne *abaa* 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).
5. G est-elle une grammaire ambiguë ? Justifiez.
6. Quelle est la probabilité de la chaîne *aaba* donnée par H (s_1 est le seul état non terminal)?
7. Quelle est la probabilité de la chaîne *aaaa* donnée par G ?

Algorithme EM (5 pts)

Dans une usine, M machines sont utilisées pour produire des pièces. Chaque machine possède une probabilité p_m de produire une pièce défectueuse. Une chaîne de production produit $T = k \times N$ pièces à l'aide de ces machines selon le procédé suivant, répété k fois de manière indépendante:

a) une machine est choisie parmi M machines possibles

b) la machine choisie est utilisée pour réaliser une série de N pièces

Vous disposez d'un relevé identifiant le nombre de pièces défectueuses de chaque série de N pièces produites. Par exemple, le relevé suivant `10, 2, 4, 5, 0` indique qu'une première série de pièces a généré 10 pièces défectueuses sur N pièces produites, la seconde série a produit 2 pièces défectueuses, etc.

1. En admettant que la génération d'une pièce défectueuse n'ait aucune influence sur la qualité des autres pièces générées par la suite, indiquez comment utiliser le relevé et l'algorithme EM pour estimer la probabilité qu'une machine soit choisie lors de la production, et la probabilité de défectuosité de chaque machine.

   Vous commencerez par formaliser le problème, vous identifierez la variable cachée et dériverez les équations de ré-estimation correspondantes.

2. Vous suspectez que dans le processus de production, l'une des machine est choisie avec une probabilité de 0.5. Indiquez une manière possible de rendre compte de cette information dans votre algorithme.