IFT6010-A2009 Int. Artif. — Trait. des Langues Naturelles **P. Langlais**
Examen Jeudi 3 décembre – 10h30 - 12h30, salle 1177

Consignes

- Vous avez le droit à vos notes de cours, livres, etc.
- Vous pouvez utilisez visualiser vos notes de cours sur votre laptop. Cependant, vous devez désactiver la connexion (wifi) à l'internet et ne faire usage d'aucune autre application que le visualisateur pdf.
- Répondez directement sur le carnet de réponse.
- Les questions appellent le plus souvent à des réponses courtes et précises.

---

(5) 1. Le coin modèle de langue

(a) Donnez la définition la plus simple et générale possible d'un modèle de langue probabiliste.

(b) Qu'est-ce qu'un modèle *n-gram* ?

(c) Quelle est la différence la plus importante entre un modèle de repli et un modèle interpolé ? Nommez un modèle de chaque vu en cours.

(d) Vous disposez d'un modèle probabiliste de la forme $P(w_{i+1}w_i|w_{i-1}w_{i-2})$. Indiquez de manière concise mais précise comment obtenir un modèle trigramme à l'aide de ce modèle.

(e) Le modèle $P(w_{i+1}w_i|w_{i-1}w_{i-2})$ est entraîné par fréquence relative sur un corpus d'entraînement. Quelle implication cela a sur votre modèle trigramme en terme de lissage ? Le cas échéant, proposez une solution pour lisser votre modèle trigramme.

(7) 2. Le coin grammaires

A) Soit la grammaire $G = \{\{S_0, S, S_1, S_2, S_3, S_4, A, B, B'\}, \{a, b\}, \mathcal{R}, S_0\}$ où $\mathcal{R}$ est:

$S_0 \rightarrow AS_1$ $S_1 \rightarrow SAB'$ $S_3 \rightarrow SA$ $A \rightarrow a$
$S_0 \rightarrow BS_2$ $S_2 \rightarrow SBB'$ $S_4 \rightarrow SB$ $B \rightarrow b$
$S \rightarrow AS_3$ $S \rightarrow BS_4$ $S \rightarrow \epsilon$ $B' \rightarrow B'B$ $B' \rightarrow B$

(a) Quel est le langage reconnu par cette grammaire ?

(b) Cette grammaire est-elle régulière ?

(c) La grammaire est-elle LL1 ? Justifiez (sans démontrer).

(d) Est-ce une bonne idée d'appliquer une technique d'analyse descendante pour analyser une phrase avec cette grammaire ? Justifiez.

B) Soit la grammaire $G' = \{\{A\}, \{a\}, \{A \rightarrow AA, A \rightarrow aa, A \rightarrow a\}, \{p_1, p_2, p_3\}, A\}$.

(a) Quel est le langage décrit par cette grammaire ?

(b) Quelle condition sur les probabilités de chaque règle (p_1, p_2, p_3) est nécessaire pour que G' soit une grammaire probabiliste ?

(c) Cette grammaire est-elle ambiguë ? Justifiez.

(d) Quelle est la probabilité de la chaîne: *aaa* ? Je ne vous demande pas de la calculer, mais de l'exprimer en fonction de p_1, p_2 et p_3.

(3) 3. Intermède

(a) Peut-on selon vous transformer un modèle bigramme en une grammaire hors-contexte probabiliste ? Le cas échéant, indiquez comment.

(3) 4. Le coin programmation dynamique

Considérez cette table d'édition entre les formes `MEALEN` et `ETAL`:

		E	T	A	L
	0	1	2	3	4
M	**1**	1	2	3	4
E	2	**1**	**2**	3	4
A	3	2	2	**2**	3
L	4	3	3	3	**2**
E	5	4	4	4	**3**
N	6	5	5	5	**4**

(a) Quelle est la distance d'édition (levenshtein distance) entre ces deux chaînes ?

(b) Quel est l'alignement induit par le chemin dont les chiffres sont en gras et encadrés ?

(c) Existe-t-il un autre alignement de ces deux chaînes de même distance ? Dans l'affirmative, indiquez lequel.

(4) 5. Le coin des modèles IBM

Considérez les deux phrases, l'une en français: f $\equiv$ *Jean aime Marie* et l'autre en anglais: e $\equiv$ *Mary is loved by Jim.*

(a) Dessinez un alignement plausible entre f et e **sous la contrainte IBM**, si le français est la langue source (modèle $P(e|f)$).

(b) Même question si la langue source est l'anglais (modèle $P(f|e)$) .

(c) Que se passe-t-il à votre avis si on tente d'aligner une phrase cible qui contient un mot qui n'a pas été vu à l'entraînement du modèle d'alignement.

(d) Soit le bitexte d'entraînement de 3 paires de phrases: $\{(db, BD), (abc, AB), (ac, CBA)\}$ où les symboles en minuscule sont des mots sources et ceux en majuscule des mots cibles. Combien de paramètres contient un modèle de transfert IBM1 ($P(cible|source)$) entraîné sur ce corpus ?

(6) 6. Le coin application

A) Vous devez gérer une application qui reconnaît des phrases courtes comme:

Je veux un billet d'avion de Paris à Montréal mardi matin,

Je voudrais réserver un billet d'avion de Paris à Montréal mercredi soir,

Réserver une chambre à Montréal mercredi,

Vous introduisez pour cela des *concepts* afin d'aider à l'analyse de ces phrases. Les phrases précédentes peuvent par exemple être étiquetées à l'aide d'un jeu de concepts comme:

[Je veux un billet d'avion]$_{avion}$ de [Paris]$_{ville}$ à [Montréal]$_{ville}$ [mardi matin]$_{time}$,

[Je voudrais réserver un billet avion]$_{avion}$ pour [Paris]$_{ville}$ depuis [Montréal]$_{ville}$ [mercredi soir]$_{time}$,

[Réserver une chambre]$_{chambre}$ à [Montréal]$_{ville}$ [mercredi]$_{time}$,

(a) Indiquez comment un modèle de markov caché peut être utilisé pour réaliser la correspondance entre une phrase et sa séquence de concepts. Je ne m'attends pas à une description verbeuse, mais plutôt une description du contenu des matrices de transition et d'émission.

(b) Étant donnée une phrase, indiquez comment à l'aide de votre modèle vous pourriez obtenir la séquence de concepts qui lui correspond. Là encore, je m'attends à une réponse concise et précise.

B) Vous souhaitez développer une application capable de transformer un texte en minuscule en un texte où les caractères devant être en majuscule le sont (première lettre de la phrase, noms propres, acronymes, etc.). Par exemple vous souhaitez transformer *paul a reçu le vaccin anti h1n1 en se rendant au clsc du plateau* en *Paul a reçu le vaccin anti H1N1 en se rendant au CLSC du Plateau.*

(a) Vous disposez d'un corpus de textes où la casse est préservée (les lettres qui doivent être en majuscule le sont). Indiquez comment utiliser un modèle HMM pour réaliser cette tâche. Vous prendrez soin d'identifier les états de votre modèle, les observations. Vous indiquerez comment entraîner les paramètres de votre modèle et préciserez comment effectuer la transformation.

(2) 7. Le coin séminaire

Résumez brièvement 2 séminaires auxquels vous avez assisté dans le cadre des séminaires RALI-OLST.

Bonne chance