Consignes

- Vous avez le droit à vos notes de cours, livres, etc.
- Vous pouvez visualiser vos notes de cours sur votre ordinateur portable. Cependant, vous devez désactiver la connexion (wifi) à l'internet et ne faire usage d'aucune autre application que votre lecteur pdf. Votre téléphone / iphone / autre gadget ne peut être utilisé que pour vous donner l'heure.
- Répondez directement sur le carnet de réponse.
- Les questions appellent le plus souvent à des réponses courtes et précises.
- Le barème (entre parenthèse) est donné à titre indicatif seulement. Le total des points à prendre dépasse les 30 points que vous aurez au maximum à cet examen.

---

(4) 1. **Le coin modèle de langue**

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

(b) Exprimez en fonction de $|u|$ et $|V|$ le dénominateur de la formule suivante pour que p définisse un modèle probabiliste bigramme sur le vocabulaire V. Justifiez.

$$p(v|u) = \frac{|uv| + \frac{1}{e^{|u|}}}{\textit{à compléter}} \quad \forall u, v \in V$$

où $|\bullet|$ est la fréquence de $\bullet$ en corpus (d'entraînement), V est l'ensemble des types de ce corpus et $|V|$ désigne la taille de V (c'est-à-dire le nombre de types).

(c) Indiquez quelle pourrait être l'intuition d'un tel modèle et en quoi il se distingue d'un modèle très proche vu en cours que vous identifierez.

(10) 2. **Le coin grammaire**

A) Soit la grammaire suivante ou S est l'axiome:

$r_1 \equiv$ Ph $\rightarrow$ Suj Verb Comp	$r_5 \equiv$ GNom $\rightarrow$ Art Adj Nom	$r_9 \equiv$ Verb $\rightarrow$ *is* | *can* | *drink*
$r_2 \equiv$ Ph $\rightarrow$ Verb Comp	$r_6 \equiv$ GNom $\rightarrow$ Art Nom	$r_{10} \equiv$ Nom $\rightarrow$ *can* | *bottle*
$r_3 \equiv$ Suj $\rightarrow$ GNom	$r_7 \equiv$ Comp $\rightarrow$ GNom	$r_{11} \equiv$ Pro $\rightarrow$ *it* | *I*
$r_4 \equiv$ Suj $\rightarrow$ Pro	$r_8 \equiv$ Adj $\rightarrow$ *empty* | *full*	$r_{12} \equiv$ Art $\rightarrow$ *the*

(a) Cette grammaire est-elle régulière ? Justifiez.

(b) Le langage décrit par cette grammaire est-il régulier ? Justifiez.

(c) Cette grammaire est-elle LL1 ? Justifiez.

(d) 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.

B) Soit $\mathcal{L}$ le langage des chaînes sur l'alphabet $\{a, b\}$ d'au moins 1 caractère qui contiennent la séquence *ab*. Les chaînes *aabb*, *aab* et *baba* appartiennent par exemple à ce langage, au contraire de *ba* et *baaa*.

(a) Écrire une grammaire qui reconnait $\mathcal{L}$ et seulement $\mathcal{L}$.

(b) Votre grammaire est-elle régulière ? Justifiez.

(c) Selon vous, $\mathcal{L}$ est-il régulier ? Justifiez.

C) Considérez le langage $\mathcal{L} \equiv \{a^n b^m \,/\, m \geq n > 0\}$.

(a) Indiquez une chaîne de $\mathcal{L}$.

(b) $\mathcal{L}$ est-il selon vous régulier ? Justifiez.

(c) Donnez une grammaire capable de reconnaître les chaînes de $\mathcal{L}$ et seulement celles-ci.

(4) 3. **Analogies**

(a) Trouvez une solution plausible à l'équation analogique:
[$\oplus$: $\bigoplus$ :: $\otimes$: ?].

(b) Trouvez une solution plausible à l'équation analogique:
[grand : géant :: gros : ?].

(c) Trouvez une solution aux équations analogiques suivantes en prenant comme définition de l'analogie formelle celle de Stroppa & Yvon vue en cours[1]:

(c_1) [déchargera : rechargerions :: déclassera : ?]
(c_2) [KaaTib : KuTaaB :: QaaRi' : ?]

(4) 4. **Le coin HMM**

Considérez le modèle markovien H dont les matrices de transition A, d'émission B et de transition initiale π sont données ci-après. On rappelle que dans un tel modèle, chaque état atteint émet une observation (ici parmi a, b et c).

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

(a) Quel est le langage reconnu par H si on admet que seul s_2 est un état final du modèle ?

(b) Quelle est la probabilité de la chaîne *aaaa* donnée par H (s_2 est le seul état final)? Je ne vous demande pas de la calculer, mais de l'exprimer en fonction des probabilités de transition et d'émission.

[1] Vous pouvez répondre sans avoir compris les détails de cette définition!

(3) 5. **Le coin IBM**

(a) Quelle est la contrainte introduite dans les modèles d'alignements IBM que l'ensemble des modèles IBM vus en cours partagent ? Pourquoi cette contrainte ?

(b) Imaginez que vous entraîniez un modèle de traduction IBM avec comme bitexte des paires de phrases identiques (la ième phrase de la partie source est identique à la ieme phrase de la cible). Quel genre d'information pensez-vous que les distributions lexicales vont capturer ? Justifiez.

(10) 6. **Le coin programmation dynamique**

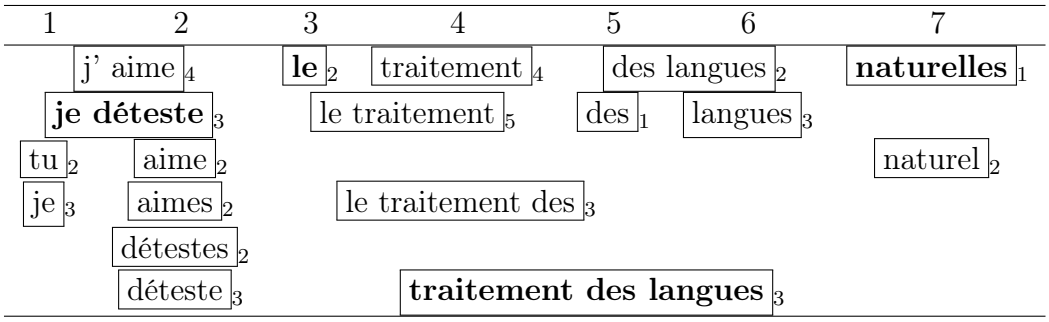
A) Génération de textes

Vous cherchez à générer des textes en juxtaposant des séquences de mots possédant un score et une position dans le texte où cette séquence peut commencer. Une séquence est donc représentée par un triplet $< w_1^n, d, s >$ où w_1^n est la séquence de mots, d la position dans le texte du premier mot de la séquence et s son score. On supposera que le score d'un texte est une fonction cumulative des scores des différentes séquences le composant et l'on cherchera ici à minimiser ce score.

Vous disposez d'un ensemble de séquences $\mathcal{E} \equiv \{< w_1^{n_i}, d_i, s_i >\}_{i \in [1,|\mathcal{E}|]}$ comme celui-ci:

$$\mathcal{E} \equiv \left\{ \begin{array}{l} < \text{j' aime}, 1, 4 >, < \text{je déteste}, 1, 3 >, < \text{tu}, 1, 2 >, < \text{je}, 1, 3 >, \\ < \text{aime}, 2, 2 >, < \text{aimes}, 2, 2 >, < \text{détestes}, 2, 2 >, < \text{déteste}, 2, 3 >, \\ < \text{traitement}, 4, 4 >, < \text{le traitement}, 3, 5 >, < \text{le}, 3, 2 >, < \text{des langues}, 5, 2 >, \\ < \text{des}, 5, 1 >, < \text{langues}, 6, 3 >, < \text{naturelles}, 7, 1 >, < \text{le traitement des}, 3, 5 > \\ < \text{traitement des langues}, 4, 3 >, < \text{naturel}, 7, 2 > \end{array} \right\}$$

que l'on peut représenter par une table où les boites représentent les séquences, le score à droite de chaque boite représente le score de la séquence, et chaque colonne indique une position dans le texte:



Les textes que l'on cherche à générer sont tous ceux que l'on peut former en juxtaposant des séquences de mots de telle façon que toutes les positions entre 1 et n (le nombre de mots du texte) soient couvertes par un mot. Ainsi *le traitement des langues* n'est pas un texte (car il commence à la position 3), tandis que *j'aime le traitement* est un texte de 4 mots (les positions 1 à 4 sont couvertes par les mots des trois séquences juxtaposées).

(a) Combien de textes peut-on former avec l'ensemble $\mathcal{E}$. Justifiez.

(b) Écrire un algorithme capable d'énumérer tous les textes possibles que l'on peut construire à partir d'un ensemble de séquences, avec leur score associé. Vous pouvez si cela vous arrange générer plusieurs fois le même texte (avec un score différent ou pas). Expliquez votre algorithme.

(c) Vous souhaitez mettre en place un algorithme de programmation dynamique pour identifier un texte de n mots de plus faible score que l'on peut former à partir d'un ensemble de séquences. Écrire une récurrence permettant de déployer une telle programmation dynamique. Expliquez cette récurrence.

(d) Indiquez comment implémenter efficacement cette récurrence. Exprimez la complexité de votre algorithme.

B) Distance d'édition

(a) Indiquez la table d'édition correspondant au calcul de la distance d'édition entre la chaine **abcd** et elle-même (la distance est donc de 0) si tous les coûts (insertion, suppression et substitution) sont unitaires.

(b) Une table d'édition peut représenter plusieurs alignements de plus faible distance. Indiquez une condition nécessaire que doit vérifier au moins une case (i, j) de la table d'édition pour qu'au moins 2 alignements de plus faible distance existent.

Bonne chance