IFT6010
Traitement Automatique des Langues Naturelles

Examen
Automne 2005
Philippe Langlais

- Inscrivez tout de suite votre nom et code permanent en bas de page
- Toute documentation 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

Nom: _____

Code Permanent: _____

Modèle de langue et lissage (5 pts)

1. À quoi sert de lisser un modèle de langue ?
2. Complétez le dénominateur de la formule suivante pour que p soit un modèle bigramme:

$$p(v|u) = \frac{|uv| + \delta(|u|)}{D(\text{à compléter})} \quad \forall u, v \in V$$

   où $|\bullet|$ est la fréquence de $\bullet$ en corpus (d'entraînement), et V est l'ensemble des types de ce corpus.
3. Donnez une expression de δ, fonction de la fréquence du contexte conditionnant en justifiant votre choix.
4. Expliquez brièvement les faiblesses de cette technique de lissage.

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:

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

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

A $\rightarrow$ *a* A `[0.1]`	A $\rightarrow$ *a* B `[0.3]`	A $\rightarrow$ *a* C `[0.6]`
B $\rightarrow$ *a* A `[0.1]`	B $\rightarrow$ *a* C `[0.1]`	B $\rightarrow$ *a* `[0.2]`
B $\rightarrow$ *b* A `[0.15]`	B $\rightarrow$ *b* C `[0.15]`	B $\rightarrow$ *b* `[0.3]`
C $\rightarrow$ *c* A `[0.8]`	C $\rightarrow$ *c* B `[0.2]`	

1. Représentez graphiquement H.
2. Quel est le langage reconnu par H si on admet que seul s_2 est un état final du modèle ?
3. Quel est le langage reconnu par G ?
4. Donnez un arbre d'analyse de la chaîne *abcb* 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 *aaaa* donnée par H (s_2 est le seul état final)?
7. Quelle est la probabilité de la chaîne *aaaa* donnée par G ?

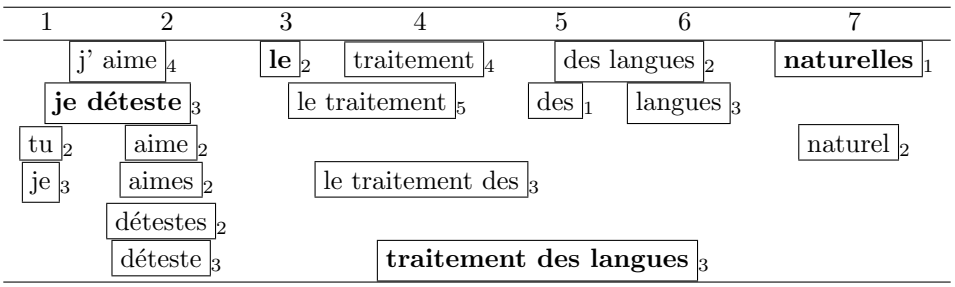
Programmation (dynamique) (10 pts)

Vous cherchez à générer un texte de N mots en juxtaposant des séquences de mots possédant un 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. Une séquence est en fait 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. Vous disposez d'un ensemble de telles séquences $S \equiv \{< w_{i1}^{n_i}, d_i, s_i >\}_{i \in [1,S]}$.

Par exemple, vous souhaitez produire des textes à l'aide de l'ensemble de séquences suivantes:

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

que l'on peut par exemple représenter comme ceci:



1. Un texte est une séquence adjacente de mots construit par juxtaposition de séquences en commencant à la position 1 (vous ne pouvez pas sauter les premières séquences) et ne contennant pas de position vide (*j'aime le traitement des langues naturelles* et *je déteste* sont des textes, pas *traitement des langues* ou *j'aime le naturel*). Énumérez tous les textes que l'on peut créer à partir de $\mathcal{E}$.

2. Écrivez un algorithme capable d'énumérer tous les textes possibles que l'on peut construire à partir d'un ensemble S 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.

3. Écrire la récurrence expliquant comment vous allez procéder pour obtenir un texte de score minimum de N mots à partir d'un ensemble S de séquences.

4. Écrivez un algorithme qui réalise cette équation. Notez que la solution consistant à générer tous les textes, puis à les trier en fonction de leur score n'est pas ce qui vous est demandé ici.

IFT6010
Traitement Automatique des Langues Naturelles

Examen
Automne 2005
Philippe Langlais

- Inscrivez tout de suite votre nom et code permanent en bas de page
- Toute documentation 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

Nom: _____

Code Permanent: _____

Answer Key for Exam A

Modèle de langue et lissage (5 pts)

1. À quoi sert de lisser un modèle de langue ?

   Answer: À distribuer une partie de la masse de probabilité des événements vus en corpus sur les événements non vus dans le but d'éviter qu'un événement non vu obtienne une probabilité nulle.

2. Complétez le dénominateur de la formule suivante pour que p soit un modèle bigramme:

$$p(v|u) = \frac{|uv| + \delta(|u|)}{D(\text{à compléter})} \quad \forall u, v \in V$$

   où $|\bullet|$ est la fréquence de $\bullet$ en corpus (d'entraînement), et V est l'ensemble des types de ce corpus.

   Answer: bla

3. Donnez une expression de δ, fonction de la fréquence du contexte conditionnant en justifiant votre choix.

   Answer: Beaucoup de réponses possibles, par exemple, on peut exprimer l'idée qu'on doit moins tricher dans les situations où le compte du contexte conditionnant est fréquent (car dans ce contexte, il la fréquence relative est un bon estimateur). Voici un exemple:
   $\delta(|u|) = \frac{1}{exp(|u|)}$)

4. Expliquez brièvement les faiblesses de cette technique de lissage.

   Answer: Si la valeur ajoutée aux comptes est la même ($\delta(|u|) = c, \forall |u|$), alors ce modèle attribue la même masse de probabilité à tous les événements non vus en corpus. Redistribue trop de masse de probabilité sur les événements non vus.

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:

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

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

A $\rightarrow$ a A `[0.1]`	A $\rightarrow$ a B `[0.3]`	A $\rightarrow$ a C `[0.6]`
B $\rightarrow$ a A `[0.1]`	B $\rightarrow$ a C `[0.1]`	B $\rightarrow$ a `[0.2]`
B $\rightarrow$ b A `[0.15]`	B $\rightarrow$ b C `[0.15]`	B $\rightarrow$ b `[0.3]`
C $\rightarrow$ c A `[0.8]`	C $\rightarrow$ c B `[0.2]`	

1. Représentez graphiquement H.

2. Quel est le langage reconnu par H si on admet que seul s_2 est un état final du modèle ?

Answer: Les chaînes sur l'alphabet $\{a, b, c\}$ d'au moins deux symboles et se terminant par a ou b et tel qu'il n'y a aucun `c` ou `b` qui se suivent.

3. Quel est le langage reconnu par G ?

   Answer: Idem.

4. Donnez un arbre d'analyse de la chaîne *abcb* 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.

   Answer: oui car il existe deux arbres d'analyse pour la chaine *aaaa*.

6. Quelle est la probabilité de la chaîne *aaaa* donnée par H (s_2 est le seul état final)?

   Answer: Il faut ici voir qu'il existe deux séquences pouvant amener à générer cette chaîne et dire que la probabilité de la chaîne est donc la somme des deux.

7. Quelle est la probabilité de la chaîne *aaaa* donnée par G ?

   Answer: idem (deux arbres). Il faut donc sommer pour obtenir la somme.

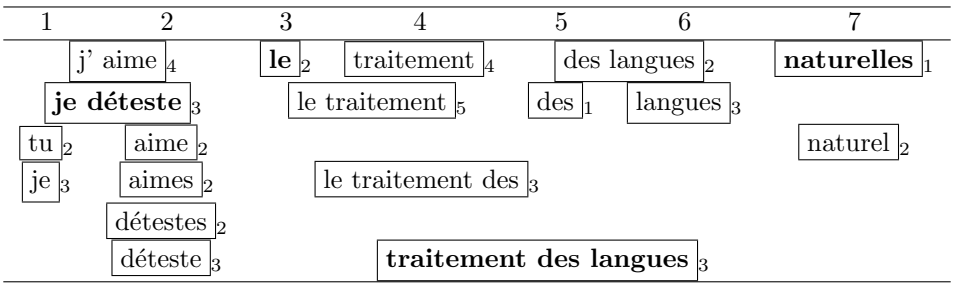
Programmation (dynamique) (10 pts)

Vous cherchez à générer un texte de N mots en juxtaposant des séquences de mots possédant un 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. Une séquence est en fait 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. Vous disposez d'un ensemble de telles séquences $S \equiv \{< w_{i1}^{n_i}, d_i, s_i >\}_{i\in[1,S]}$.

Par exemple, vous souhaitez produire des textes à l'aide de l'ensemble de séquences suivantes:

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

que l'on peut par exemple représenter comme ceci:



1. Un texte est une séquence adjacente de mots construit par juxtaposition de séquences en commencant à la position 1 (vous ne pouvez pas sauter les premières séquences) et ne contennant pas de position vide (*j'aime le traitement des langues naturelles* et *je déteste* sont des textes, pas *traitement des langues* ou *j'aime le naturel*). Énumérez tous les textes que l'on peut créer à partir de $\mathcal{E}$.

   Answer:

2. Écrivez un algorithme capable d'énumérer tous les textes possibles que l'on peut construire à partir d'un ensemble S 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.

   Answer:

   ```
   enumere(texte,sc,deb):
   for s débutant en deb do
     enumere(texte+words(s),sc+score(s),deb+pos(s)))
     print(texte,sc)
   ```

3. Écrire la récurrence expliquant comment vous allez procéder pour obtenir un texte de score minimum de N mots à partir d'un ensemble S de séquences.

   Answer:

   $$M[a,b] = \left\{ \begin{array}{l} \min_{c\in[a,b[} M[a,c] + M[c+1,b] \\ \min_{s_i\in S: pos(s)=a \text{ et } n_i=(b-a+1)} score(s_i) \end{array} \right.$$

4. Écrivez un algorithme qui réalise cette équation. Notez que la solution consistant à générer tous les textes, puis à les trier en fonction de leur score n'est pas ce qui vous est demandé ici.

   Answer: C'est comme un CYK. On génère toutes les séquences de taille n en ordre croissant de n. On a pour cela besoin d'une table à 2 dimensions $T[d,f]$ indiquant le texte de plus faible coût s'étendant sur ces positions.