Programmation d'un étiqueteur morpho-syntaxique

à rendre pour le XX

L'étiquetage morpho-syntaxique consiste à associer chaque mot d'une séquence sa catégorie correcte (verbe, nom, adjectif, adverbe ...). La complexité du problème provient du fait que certains mots peuvent avoir plusieurs catégories selon leur position dans la phrase (... **la** maison ..., ... il **la** donne ..., ... le **la** est une note de musique ...).

L'objectif de ce projet est d'implémenter un étiqueteur morpho-syntaxique à l'aide d'un HMM. Dans sa version la plus simple, le HMM possède autant d'états qu'il existe de catégories différentes. Les observables sont constitués par les mots.

Pour estimer les paramètres du HMM, on dispose de données complètes : une collection de phrases dont chaque mot a été associé à sa catégorie correcte. On peut alors estimer les paramètres du HMM par fréquence relative :

– Les probabilités initiales :

$$\pi(c) = P(X_1 = c) \approx \frac{\mathcal{I}(c)}{\mathcal{N}}$$

où $\mathcal{I}(c)$ est le nombre de phrases commençant par la catégorie c et $\mathcal{N}$ est le nombre de phrases.

– Les probabilités de transition :

$$T(c_1, c_2) = p(X_t = c_2 | X_{t-1} = c_1) \approx \frac{\mathcal{C}(c_1, c_2)}{\mathcal{C}(c_1)}$$

où $\mathcal{C}(c_1, c_2)$ est le nombre d'occurrences du bigramme c_1c_2 et $\mathcal{C}(c_1)$ le nombre d'occurrences de la catégorie c_1.

– Les probabilités d'émission :

$$E(c, m) = P(O_t = m | X_t = c) \approx \frac{\mathcal{C}(m, c)}{\mathcal{C}(c)}$$

où $\mathcal{C}(m, c)$ est le nombre de fois que m a été étiqueté c et $\mathcal{C}(c)$ le nombre d'occurrences de la catégorie c.

Les données d'apprentissage se trouvent dans le fichier `train`. Dans le fichier `test`, on trouvera d'autres phrases étiquetées qui serviront à calculer les performances de l'étiqueteur. Les phrases de test ne seront pas utilisées pour estimer les paramètres, elles sont sensées représenter des nouvelles phrases.

La mesure utilisée pour l'évaluation de l'étiqueteur est la précision, qui mesure tout simplement la proportion de mots du corpus de test auxquelles l'étiqueteur a

attribué la bonne catégorie.

Vous pourrez utiliser l'implémentation de HMM qui se trouve dans le fichier `hmm.c`. Elle permet de représenter un HMM sous la forme de trois tableaux, un tableaux pour les probabilités initiales un tableau pour les probabilités de transitions et un tableau pour les probabilités d'émission.

Dans l'implémentation proposée, un HMM peut être stocké dans un fichier texte au format suivant [1] :

```
#nb etats
2
#nb observables
2
#probabilites initiales
0.362146 # I(0) = P(X_1 = 0)
0.637854 # I(1) = P(X_1 = 1)
#probabilites de transition
0.479827 # T(0,0) = P(X_i = 0 | X_{i-1} = 0)
0.520173 # T(0,1) = P(X_i = 1 | X_{i-1} = 0)
0.819147 # T(1,0) = P(X_i = 0 | X_{i-1} = 1)
0.180853 # T(1,1) = P(X_i = 1 | X_{i-1} = 1)
#probabilites d'emission
0.257264 # E(0,0) = P(0_i = 0 | X_i = 0)
0.742736 # E(0,1) = P(0_i = 1 | X_i = 0)
0.742653 # E(1,0) = P(0_i = 0 | X_i = 1)
0.257347 # E(1,1) = P(0_i = 1 | X_i = 1)
```

L'implémentation qui vous est donnée ne manipule que des entiers (les états et les observables sont identifiés par des entiers successifs à partir de zéro). Il est donc nécessaire d'encoder les données avant de pouvoir construire le HMM correspondant à l'étiqueteur. Pour cela deux fichiers vous sont fournis : le fichier `voc_observables` qui contient la liste des mots présents dans les données d'entrainement et de test et le fichier `voc_etats` qui contient la liste des catégories. On considèrera que le premier mot du fichier `voc_observables` est associé à l'entier 0, le second à l'entier 1 et ainsi de suite. On suivra le même principe pour les catégories.

Votre travail consiste donc à :

1. écrire un programme d'encodage et de décodage des données
2. écrire un programme qui prend en entrée le corpus de test encodé et produit un HMM sous la forme d'un fichier texte tel que décrit au dessus
3. programmer l'algorithme de Viterbi qui permettra de retrouver pour une séquence d'observables (de mots) la séquence d'états (de catégories) la plus probable
4. calculer la précision de votre étiqueteur sur les données de test

1. Les caractères qui suivent un dièse sont des commentaires, ils peuvent être omis.