TP2 - Analyse morphologique

Programmation pour le TAL - SDL & SCA

4 mars 2013

Pour chacun des exercices, sauvegardez le script Python dans un fichier qui a comme nom *exo <numéro de l'exercice>.py* puis mettre tous les fichiers dans une archive .zip qui a comme nom *TP2_<nom de l'étudiant>_<prénom de l'étudiant >.zip*. A la fin de la séance, envoyez l'archive en fichier attaché à l'adresse *perrier@loria.fr* en mettant comme objet du message *programmation pour le TAL : TP2*.

1 Approche naïve sans utilisation d'un lexique

Faire l'analyse morphologique d'un texte segmenté, c'est pour chacun des mots du texte, retrouver le lemme correspondant, sa catégorie grammaticale ainsi que les paramètres de flexion. En voici quelques exemples :

mot	**lemme**	**catégorie**	**paramètres de flexion**
maisons	maison	nom	nombre : pluriel
chevaux	cheval	nom	nombre : pluriel
infirmière	infirmier	nom	nombre : singulier, genre : féminin
blanche	blanc	adjectif	nombre : singulier, genre : féminin
fragiles	fragile	adjectif	nombre : pluriel
délicieux	délicieux	adjectif	genre : masculin
mangeais	manger	verbe	mode : indicatif, temps : imparfait, nombre : singulier, personne : 1|2
va	aller	verbe	mode : indicatif, temps : présent, nombre : singulier, personne : 3
repris	prendre	verbe	mode : participe, temps : passé, genre : masculin
ferme	ferme	nom	nombre : singulier
ferme	ferme	adjectif	nombre : singulier
ferme	fermer	verbe	mode : indicatif, temps : présent, nombre : singulier, personne : 1|3

Exercice 1.1 *Ecrivez un script Python prenant en entrée un nom sous forme d'une chaîne de caractères et fournissant en sortie son lemme, son nombre et éventuellement son genre (pour les noms ayant un masculin et un féminin). Vous vous servirez de la forme des suffixes des mots et des règles de flexion du nom en français, règles que vous pourrez trouver sur le Web ; par exemple, vous pouvez aller à l'adresse suivante : http ://fr.wikipedia.org/wiki/Morphologie_du_nom_en_français.*

En commentaires, vous expliquerez à la fin du programme les limites d'une telle approche.

Exercice 1.2 *Faites la même chose qu'à l'exercice précédent pour les adjectifs avec les mêmes paramètres, apparaissant seulement quand ils entraînent une modification de la flexion du mot.*

Exercice 1.3 *Ecrivez un script Python qui prenne en entrée un verbe conjugué à l'indicatif présent du premier groupe (verbe qui se termine par* er *à l'infinitif) et qui retourne en sortie son lemme, c'est-à-dire le verbe à l'infinitif, ainsi que la valeur de ses paramètres de flexion : nombre, personne.*

Pour cela, vous utiliserez le fait que la forme conjuguée d'un verbe du premier groupe s'obtient par ajout d'un suffixe à la racine selon des règles très précises. Par exemple, mang + ent *donne* mangent.

Vous ferez attention au fait que dans certains cas très délimités, la racine peut subir des variations mineures : lever, payer, jeter, ennuyer ... *Vous pourrez trouver toute l'information à ce sujet sur le Web.*

Exercice 1.4 *Ecrivez un script Python qui fasse l'analyse morphologique des noms, adjectifs, verbes du premier groupe à l'indicatif présent repèrés dans un texte.*

Plus précisément, l'entrée du programme sera un texte où les noms sont immédiatement suivis de la balise < N >*, les adjectifs de la balise* < A > *et les verbes du premier groupe à l'indicatif présent par la balise* < V >*.*
Voici un exemple d'entrée : Les petites<A> filles<N> jouent<V> avec les chevaux<N>.

En sortie, on veut un texte où les mots étiquetés ont été remplacés par leur lemme, accompagné de sa catégorie grammaticale et des paramètres de flexion.

Pour l'exemple précédent, cela donnerait : Les <petit, N, nb=pl, gen=f> <fille, N, nb=pl> <jouer, V, mode=ind, temps=pres, nb=pl, pers=3> avec les <cheval, N, nb=pl>.

A la fin du programme, indiquez en commentaires les points forts et les points faibles d'une telle approche de l'analyse morphologique.

2 Utilisation d'un lexique

Il existe une autre approche qui exige un lexique morphologique, c'est-à-dire un lexique des mots de la langue, où, pour chaque mot, on trouve son lemme, sa catégorie grammaticale et ses paramètres de flexion. Voici un exemple réduit d'un tel lexique.

blanche	<blanc, adjectif, nb=sg, gen=f >
dans	<dans, préposition >
ferme	<ferme, nom, nb=sg >
ferme	<ferme, adjectif, nb=sg >
ferme	<fermer, verbe, mode=ind, temps=pres, nb=sg, pers=1|3>
infirmière	< infirmier, nom, nb=sg, gen=f >
les	<f le, déterminant, nb=pl, gen=m>
mangeais	< manger, verbe, mode=ind, temps=imp, nb=sg, pers=1|2>
qui	<qui, pronom >
repris	<prendre, verbe, mode=part, ,temps=passe, nb=sg, gen=m>
souvent	<souvent, adverbe >

Exercice 2.1 *Ecrivez un script Python qui fasse l'analyse morphologique d'un texte en utilisant un lexique morphologique.*

Vous commencerez par écrire un lexique d'une douzaine d'entrées sous forme d'une liste. Chaque entrée sera elle-même représentée par une liste. Par exemple, voici l'entrée pour jouent *:* ['jouent', 'jouer', 'mode=ind', 'temps=pres', 'nb=pl', 'pers=3'].

Le lexique est ordonné selon l'ordre lexicographique des mots fléchis.

Après avoir écrit le lexique, vous écrirez le script qui prend en entrée un texte brut et qui fournit en sortie un texte où les mots ont été remplacés par les entrées correspondantes du lexique. Si le mot rencontré n'est pas dans le lexique, à la place du lemme, sera mis 'mot inconnu'.

A la fin du script, vous indiquerez en commentaires les points forts et les points faibles de cette approche.

Exercice 2.2 *Avec le même lexique qu'à l'exercice précédent, vous écrirez un script qui fait la tâche inverse : il prend en entrée la sortie du programme précédent mais sans ambiguïté et sans mots inconnus et il retourne l'entrée.*