option LTAL

Linguistique et Traitements Automatiques des Langues

TP 3

Un exemple d'exploitation de propriétés très générales des langues :

Diagnostic automatique de la langue (majoritaire) d'un document

Spécifications :

Le programme à réaliser prend en entrée un fichier html (par exemple articles de journaux téléchargés de sites d'organes de presse) et en diagnostique la langue majoritaire, en faisant l'hypothèse que ce document est monolingue, ou qu'une langue y est très majoritaire.

Les ressources seront constituées pour 6 langues : français, anglais, allemand, néerlandais, espagnol, grec.

Ce diagnostic sera associé à une évaluation de la qualité du diagnostic, en particulier dans 2 cas :

- 2 langues proches pour lesquelles on a des ressources (allemand - néerlandais),
- 2 langues proches, une pour laquelle on a des ressources, et une pour laquelle on n'en a pas (allemand - suédois, ou néerlandais - suédois, ou espagnol - portugais).

Les tests seront donc faits sur 8 langues : français, anglais, allemand, néerlandais, espagnol, grec, ET suédois, portugais.

Propriété linguistique très générale exploitée par le diagnostiqueur :

Les mots très fréquents et très courts sont caractéristiques d'une langue (cf. Zipf).

Méthode de réalisation :

1) Constituer le corpus d'étude
2) Valider la propriété sur le corpus d'étude
3) Constituer automatiquement les ressources du diagnostiqueur
4) Réaliser le diagnostiqueur proprement dit
5) Constituer le corpus de test et réaliser les tests sur le corpus de test

Constituer automatiquement les ressources du diagnostiqueur

À partir du corpus d'étude (6 articles pas trop courts en 6 langues), extraire automatiquement les mots très fréquents et très courts

À vous d'expérimenter pour voir si les 2 propriétés de fréquence et de longueur sont nécessaires, ou si une seule suffit (vous savez que fréquence et longueur sont corrélées).

Quel nombre de mots ? Entre 10 et 50. À vous d'expérimenter pour voir si ce critère est important.

La propriété vous semble-t-elle validée sur ce corpus ?

Réaliser le diagnostiqueur proprement dit

En voici l'algorithme général :

à partir du fichier extérieur des ressources, charger entièrement les ressources des 6 langues en RAM
charger entièrement le source html en RAM dans une chaîne unique
prétraiter le source html (enlever les javascript et les commentaires)
reconnaître l'encoding original, et décoder dans l'encoding interne python
translittérer les entités SGML
compter le nombre de mots des ressources, pour chacune des 6 langues, d'où le score de chaque langue
sortir les scores des 6 langues par ordre décroissant des scores
la langue de score maximal est considérée comme étant la langue du document
qualité du diagnostic :
- diagnostic sûr si la langue du 2e score est suffisamment plus faible que la langue de score maximal (moins de la moitié, par exemple)
- diagnostic incertain dans le cas contraire
sortir dans un fichier html les scores de chaque langue, le diagnostic et sa qualité

Question : comment détecter qu'on n'a pas les ressources de la langue du document ?

Réaliser les tests

Constituer un corpus de test : 8 articles de presse des 8 langues à tester (corpus différent du corpus d'étude qui a servi à constituer les ressources et valider la propriété).

Placer le diagnostiqueur dans une fonction
Écrire un programme de test qui diagnostique la langue de chacun des 8 articles
Faire un compte-rendu des tests.

Extension à d'autres langues

Dans le TP 1-2, vous avez évalué la variété des pratiques du mot dans les écritures de langues variées. Comment se comportera ce diagnostiqueur de langues pour des langues ayant une pratique différente du mot ? Faites un test sur le finnois.

Travail à remettre :

Travail en binôme, à remettre en 1 exemplaire par binôme, au plus tard le **20 ou le 22 octobre** avant votre TP
Chaque binôme envoie un mail à <Julien.Gosme@info.unicaen.fr> avec le fichier du rapport, le programme réalisé, les fichiers placés en entrée et ceux obtenus en sortie.
Le rapport devra comporter :

- une courte introduction présentant le sujet du TP;
- partie 1 : compte-rendu de la constitution des ressources et de la validation de la propriété ;
- partie 2 : compte-rendu de la réalisation du diagnostiqueur;
- partie 3 : compte-rendu des tests ;
- une conclusion rappelant l'ensemble des problèmes rencontrés et les solutions trouvées pour les résoudre;
- une annexe avec les pages de **résultats commentés** ;

(nombre de pages limité à 5, annexes non comprises; soin particulier pour la rédaction et l'orthographe).

•