ALOUI Cindy
DARY Franck

M1 Informatique

Cahier des charges :
Correction de texte "chat" par approche de traduction automatique.

Contexte et définition :

Dans le traitement automatique du langage naturel (TALN), il convient d'effectuer sur les données un traitement de lemmatisation pour se débarrasser des chaînes de caractères et travailler sur des unités atomiques - les lexèmes - qui facilitent grandement tout traitement.

A notre époque, l'écrasante majorité des documents en langage naturel proviennent d'internet. Qu'il s'agisse de messages de discussion instantanée (chat service client Orange) ou de publications sur les réseaux sociaux (Tweeter), dans ce type de communication le sens prime sur la forme et il est très courant que les mots ne soient pas correctement orthographiés. Cela ne nuit en rien à la communication humain/humain, mais cela rend inefficace l'étape de lemmatisation susmentionnée.

Objectif :

L'objectif de ce projet est de proposer un programme informatique qui serait capable de corriger un grand nombre d'erreurs dans n'importe quel texte en langage naturel , qui pourra ainsi être utilisé pour prétraiter un corpus sur lequel on souhaite ensuite appliquer un outil de TALN.

Limites :

Le programme ne fonctionnera qu'avec des textes en français.
La syntaxe et la grammaire sont beaucoup plus difficiles à corriger que l'orthographe, mais sont aussi moins importants. Le plus important étant qu'à un mot soit associé le bon lexème, peu importe le genre ou le nombre. C'est pourquoi notre programme se contentera de corrections allant dans ce sens et ne modifiera pas la syntaxe du texte. Ainsi, un texte en français correct désignera un texte dans lequel les mots sont correctement lemmatisables.

Description fonctionnelle :

Le programme a besoin pour fonctionner d'un corpus "traduit" texte chat vers texte en français correct. A partir ce corpus le programme pourra générer une table de traduction.

Le programme pourra aussi - et c'est sa fonction première - générer un texte en français correct à partir d'un texte en français incorrect.