Master 2 ATAL

Applications Multiligues

Segmentation en mots du Japonais

Étudiant :
Coraline Marie

Encadrant :
Florian Boudin

3 novembre 2014

UNIVERSITÉ DE NANTES

Table des matières

Introduction

Le Japonais est une langue très particulière, dont l'écriture diffère totalement des langues latines. L'utilisation d'alphabets différents et l'absence de marqueurs explicites, tel que les espaces, la rende plus délicate à traiter, notamment dans le cadre de la Recherche d'Information.

Afin de faciliter le traitement de textes en langue Japonaise, il faut au préalable appliquer un processus de segmentation en mots, appelée *tokenisation*. Ce rapport présente donc différentes méthodes combinables permettant de segmenter en mots, des textes en japonais. Il y aura d'abord une desription de l'implémentation de la méthode des Modèles de Markov de Constantine P. Papageorgiou[1]. Il y aura également une présentation de l'amélioration de cette méthode par l'implémentation de ngrammes, avant de finir par l'explication d'une méthode reposant sur l'analyse des différents alphabets utilisés par la langue Japonaise.

Première approche

Implémentation des trigrammes

Reconaissance des alphabets

Comme il l'a été dit précédemment, le Japonais est composé de plusieurs alphabet :

Conclusion

Bien que les scores obtenus grâce à ses diverses méthodes soient satisfaisants (94.6% de f-mesure sur le corpus de test), il reste encore des erreurs. Ceci peut être repproché au corpus d'entrainement dont la taille reste limitée, mais pas seulement.

Il y a eu pendant ce projet d'autres idées qui n'ont pas eu le temps d'être implémentées, comme par exemple :

- le traitement des alphabets par ngrammes ;
- l'ajout d'un dictionnaire : ce qui devrait éliminer les groupes de caractères inexistants ;
- l'entrainement du modèle de Markov caché sur d'autres types de corpus tokenisés :
  - un corpus web japonais (ex : Wikipédia) avec récupération des *liens* : ce qui permettrait de mettre en évidence des mots tokenisés dans leur contexte ;
  - un corpus journalistique (ex : blogs de presse, journaux, ...) avec reconnaissance de formes et de mise en page (titres, texte en italique, texte en gras, ...) ce qui pourrait également aider à la reconnaissance des mots tokenisés.

Bibliographie

[1] Constantine P. Papageorgiou. Japanese word segmentation by hidden markov model. *Proceedings of the workshop on Human Language Technology*, HLT '94, 1994.