Segmentation en mots du Japonais

Florian Boudin

Recherche d'information cross-lingue - 2013

1 Introduction

La langue Japonaise possède de nombreuses caractéristiques qui la rend (plus ?) difficile à traiter du point de vue de la Recherche d'Information (RI). L'absence de marqueurs explicites (e.g. espaces) pour délimiter les mots en fait partie. Afin de pouvoir indexer les documents rédigés en langue japonaise, un processus de segmentation en mots (*tokenisation*) doit être au préalable appliqué sur chaque document. Le découpage d'une phrase en mots peut être ambigu du fait de l'absence de marqueurs entre ces derniers. Les erreurs de segmentation sont donc fréquentes et elles ont un impact direct sur la précisions des systèmes de RI.

2 Travail demandé

Votre tâche consiste à développer un système de segmentation en mots pour le Japonais. Pour cela, vous disposez d'un ensemble d'apprentissage composé de 4000 phrases. Chaque phrase de cet ensemble a été segmentée manuellement, un exemple est montré ci-dessous :

ここで私は２つの選択肢を迫られました。
ここ　で　私　は　２　つ　の　選択肢　を　迫ら　れ　ました　。

Dans un premier temps, vous devez implémenter la méthode basée sur un modèle de Markov caché (HMM, *Hidden Markov Model*) présentée dans [1] (lien direct : `http://acl.ldc.upenn.edu/H/H94/H94-1054.pdf`). Bien que simple, cette méthode donne de bons résultats (89.3% de f-mesure sur le corpus de test).

Dans un second temps, réfléchissez aux moyens d'améliorer cette méthode. Proposez et implémentez vos idées d'amélioration dont vous vérifierez ensuite l'impact sur l'ensemble de test qui vous est fourni.

Références

[1] Constantine P. Papageorgiou, *Japanese word segmentation by hidden Markov model*. Proceedings of the workshop on Human Language Technology (HLT '94), pages 283–288, 1994.