Segmentation en mots du Japonais

Florian Boudin

Recherche d'information cross-lingue - 2013

1 Introduction

La langue Japonaise possède de nombreuses caractéristiques qui la rendent 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 ont un impact important sur la précisions des systèmes de RI.

2 Travail demandé

Votre tâche consiste à développer un système supervisé de segmentation en mots pour le Japonais. Vous disposez d'un corpus d'entraînement composé de 4000 phrases segmentées manuellement.