Utilisation de WordNet dans la catégorisation de textes multilingues

Mohamed Amine Bentaallah*,** Mimoun Malki*,***

*Département d'informatique, Université Djillali Liabès, 22000 Sidi Bel Abbès, ALGERIE
http://www.univ-sba.dz
**mabentaallah@univ-sba.dz
***malki-m@yahoo.com

Résumé. Cet article est consacré au problème de la catégorisation multilingue qui consiste à catégoriser des documents de différentes langues en utilisant le même classifieur. L'approche que nous proposons est basée sur l'idée d'étendre l'utilisation de WordNet dans la catégorisation monolingue vers la catégorisation multilingue.

1 Introduction

La Catégorisation de Textes (C.T) consiste à assigner une ou plusieurs catégories parmi une liste prédéfinie à un document. En d'autres termes, elle permet de chercher une liaison fonctionnelle entre un ensemble de textes et un ensemble de catégories (Sebastiani (2002)). La grande importance accordée cette dernière décennie au traitement des données multilingues, a donné naissance à un nouveau domaine de recherche. C'est la catégorisation de textes multilingues.

Dans cet article, nous allons proposer une nouvelle approche qui consiste à étendre l'utilisation de WordNet en C.T pour catégoriser des documents provenant de différentes langues. L'approche proposée est basée sur la traduction des documents à catégoriser vers la langue de Shakespeare afin de pouvoir bénéficier de l'utilisation de WordNet par la suite. Cette hybridation entre l'utilisation des techniques de traduction et l'utilisation de WordNet offre les avantages suivants:

– Sans l'utilisation des techniques de traduction, il devient nécessaire de construire une ontologie WordNet pour chaque langue. Cette construction est très coûteuse en terme de temps et personnels.

– L'utilisation d'une ontologie bien construite et riche tel que WordNet permet de corriger certains erreurs de traduction en utilisant des relations tel que l'hypéronymie et la synonymie(Cruse (1986)).