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)).

2 L'approche proposée

L'approche que nous proposons dans cet article se compose de deux phases. La première phase est la phase d'apprentissage, elle consiste à:

- Utiliser WordNet pour mapper les mots en synsets;
- Enrichir l'espace de représentation par l'extraction des hypéronymes ;
- Utiliser la méthode χ_2 multivariée (Clech et al. (2003)) pour sélectionner les synsets caractérisant chaque catégories par rapport aux autres catégories. Les synsets sélectionnés pour chaque catégorie forment son profil conceptuel.

La deuxième phase est celle de la classification, elle consiste à assigner le texte à catégoriser à une ou plusieurs catégories en se basant sur les profils conceptuels déjà trouvés dans la première phase. Cela nécessite les étapes suivantes:

- Traduire le texte à catégoriser vers la langue anglaise afin de pouvoir utiliser WordNet pour créer le vecteur conceptuel ;
- Pondérer les profils conceptuels des catégories ainsi que le vecteur conceptuel du texte à catégoriser ;
- Calculer la distance entre le vecteur conceptuel du texte à catégoriser avec les profils conceptuels des catégories ;

3 Expérimentations

Afin de pouvoir montrer l'utilité de l'utilisation de WordNet en catégorisation multilingue, nous avons testé l'approche proposée sur le corpus monolingue Reuters21578 [1], ainsi que sur le corpus bilingue ILO. Les résultats des expérimentations ont montré que les performances obtenues sur les deux corpus sont très proches. Ce rapprochement dans les résulats nous mènent à confirmer que l'utilisation de WordNet en catégorisation multilingue est une piste prometteuse.

Références

Clech, J., R. Rakotomalala, et R. Jalam (2003). Sélection multivariée de termes. *In SFDS03*.

Cruse, D.-A. (1986). Lexical semantics. *Cambridge University Press*.

Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM Computing Surveys*, 1–47.

Summary

This article is essentially dedicated to the problem of Multilingual Text Categorization, that consists in classifying documents in different languages according to the same classification tree. The proposed approach is based on the idea to spread the utilization of WordNet in Text Categorization towards Multilingual Text Categorization.

1. disponible sur http://www.daviddlewis.com/resources/testcollections/reuters21578.