EDA : algorithme de désuffixation du langage médical[1]

Didier Nakache*,**, Elisabeth Metais*, Annabelle Dierstein*

*CEDRIC CNAM: 292 rue Saint Martin - 75003 Paris, France
**CRAMIF: 17 / 19 rue de Flandre - 75019 Paris, France
datamining@wanadoo.fr , metais@cnam.fr

1 Introduction

Pour améliorer l'efficacité des algorithmes de classification, il existe plusieurs algorithmes de préparation des données, dont la désuffixation. Cependant, le langage médical, et les comptes rendus hospitaliers sont rédigés dans un langage très technique, avec peu de formes flexionnelles. Nous nous sommes demandés si l'implémentation d'un algorithme de désuffixation dans ce contexte pouvait améliorer significativement les résultats obtenus. Nous avons mis en évidence qu'il était possible d'obtenir de meilleurs résultats que les algorithmes actuels d'une part en développant un algorithme spécifique basé sur un large corpus de documents, d'autre part en enrichissant ces derniers en fonction des racines lexicales des termes médicaux.

Plusieurs algorithmes de désuffixation ont été proposés, les plus célèbres d'entre eux étant Porter (1980), Lovins (1968) et Paice (1996). Malheureusement, il s'agit d'algorithmes de désuffixation pour la langue anglaise, dont les dérivés morphologiques se prêtent facilement à ce type d'adaptation.

2 Présentation de l'algorithme EDA et résultats

Afin d'améliorer les performances des algorithmes de classification de comptes rendus hospitaliers (projet Rhea), nous proposons une technique de désuffixation qui donne des résultats intéressants dans le contexte médical. Nous nous sommes constitué une base de 29 393 comptes rendus, tous utilisés dans cette étude. Par ailleurs, la terminologie médicale possède une structure sémantique forte. Jujols (1991).

L'algorithme EDA fonctionne en deux phases. La première phase consiste à préparer le mot en appliquant quelques modifications (transformation en minuscules, séparation des caractères ligaturés, suppression des signes diacritiques, etc.). La seconde phase consiste à enrichir le corpus de textes en fonction des structures sémantiques des termes (par exemple : foie=hépat, langue=glosso, rate=spléno, cœur=cardio,…).

[1] Ce travail a été partiellement financé par le MENRT dans le cadre du projet RNTS Rhéa.

Pour expérimenter nos résultats, nous avons choisi d'utiliser Naïve Bayes comme algorithme de classification, et la F-mesure pour l'évaluation. Ce qui donne les résultats suivants :

Désuffixation	Résultat (F-mesure)
Aucune désuffixation	69.23%
Désuffixation avec Carry	72.27%
Désuffixation avec EDA	74.72%

TAB. 1 – *Gains sur la F-mesure selon la méthode utilisée*.

3 Conclusion et perspectives

Sur 25 275 termes différents présents dans 30 000 comptes rendus, 10 602 ont été regroupés, soit 42%. L'utilisation de cet algorithme de désuffixation nous a permis de mesurer une amélioration de 5.49 %. Les deux tiers du gain résultent de la désuffixation, le dernier tiers de l'enrichissement des documents par la recherche de racines lexicales des termes médicaux.

Références

Jujols P, Aubas P, Baylon C et al. (1991) Morphosemantic Analysis and Translation of Medical Compound Terms. *Meth Inform Med*; 30:30-5.

Lovins J.B. (1968) Development of a Stemming Algorithm, *Mechanical Translation and Computational Linguistics*, 11 (1-2), 22-31.

Paice C. (1996) Method for evaluation of stemming algorithms based on error counting, *Journal of the American Society for Information Science.*

Porter M. (1980) An algorithm for suffix stripping, *Program*, 14 (3), 130-137.

Summary

Desuffixing is an easy technique for textual data processing. We apply it to French medical report for automatic classification. This paper proposes a new desuffixer algorithm adapted to the medical language.