Choix du taux d'élagage pour l'extraction de la terminologie. Une approche fondée sur les courbes ROC

Mathieu Roche*, Yves Kodratoff**

*LIRMM - UMR 5506, Université Montpellier 2,
34392 Montpellier Cedex 5 - France
mathieu.roche@lirmm.fr
**LRI - UMR 8623, Université Paris-Sud,
91405 Orsay Cedex - France
yk@lri.fr

Résumé. Le choix du taux d'élagage est crucial dans le but d'acquérir une terminologie de qualité à partir de corpus de spécialité. Cet article présente une étude expérimentale consistant à déterminer le taux d'élagage le plus adapté. Plusieurs mesures d'évaluation peuvent être utilisées pour déterminer ce taux tels que la précision, le rappel et le F_{score}. Cette étude s'appuie sur une autre mesure d'évaluation qui semble particulièrement bien adaptée pour l'extraction de la terminologie : les courbes ROC (Receiver Operating Characteristics).

1 Introduction

Cet article présente une étude expérimentale consistant à évaluer le taux d'élagage le plus adapté pour l'extraction de la terminologie. Nous allons décrire ci-dessous notre méthode globale d'extraction de la terminologie et rigoureusement définir l'élagage.

La première phase de notre travail d'extraction de la terminologie à partir de corpus spécialisés consiste à normaliser les textes en utilisant des règles de nettoyage décrites par Roche (2004). Les corpus que nous utilisons sont décrits dans la section 3 de cet article. L'étape suivante consiste à apposer des étiquettes grammaticales à chacun des mots du corpus en utilisant l'étiqueteur ETIQ développé par Amrani et al. (2004). ETIQ est un système interactif s'appuyant sur l'étiqueteur de Brill (1994) qui améliore la qualité de l'étiquetage de corpus spécialisés. Nous pouvons alors extraire l'ensemble des collocations Nom-Nom, Adjectif-Nom, Nom-Adjectif[1], Nom-Préposition-Nom d'un corpus spécialisé. L'étape suivante consiste à sélectionner les collocations les plus pertinentes selon des mesures statistiques décrites par Roche et al. (2004c); Roche (2004). Les collocations sont des groupes de mots définis par Halliday (1976); Smadja (1993). Nous appelons *termes*, les collocations pertinentes.

Les termes binaires (ou ternaires pour les termes prépositionnels) extraits à chaque itération sont réintroduits dans le corpus avec des traits d'union afin qu'ils soient reconnus comme des mots à part entière. Nous pouvons ainsi effectuer une nouvelle recherche terminologique à partir du corpus avec prise en compte de la terminologie du domaine acquise aux étapes précédentes. Notre méthode itérative, proche des travaux de Evans et Zhai (1996), est décrite

[1] Corpus en français uniquement