Des textes aux associations entre les concepts qu'ils contiennent

Yves Kodratoff*, Jérôme Azé*,
Mathieu Roche*, Oriane Matte-Tailliez* **

* CNRS, LRI ** CNRS, IGM
Université Paris Sud, 91405 Orsay Cedex
{yk,aze,roche,oriane}@lri.fr

Résumé. Nous présentons dans cet article, une chaîne originale d'outils allant de l'acquisition du corpus à l'extraction d'information. Ces outils permettent de faciliter le travail de l'expert en automatisant une partie des traitements. Nous étudions l'automatisation d'une étape clef préalable à la construction d'une ontologie terminologique, à savoir l'acquisition des termes pertinents qui constitueront les noeuds de l'ontologie. Nous avons obtenu la terminologie complète de quatre corpus différents par la langue et par la taille. La validation de ces terminologies par des experts montre que notre méthode fournit un très grand nombre de termes de qualité satisfaisante. Des classes de concepts ont été construites avec ces termes de façon semi-automatique. Celles-ci nous permettent de représenter chaque corpus sous une forme plus compacte, à partir desquelles un processus d'extraction de règles d'association peut être appliqué. Nous avons validé les règles d'association obtenues en comparant nos résultats avec ceux d'une amélioration récente de l'Intensité d'Implication sur trois corpus. Deux de ces corpus sont issus de données réelles et un expert du domaine a discuté l'intérêt des règles obtenues avec les deux mesures.

1 Introduction

Nous présentons une approche originale permettant d'extraire des connaissances à partir de corpus spécialisés. La description des quatre corpus étudiés est abordée dans la section 2.1.

Pour traiter la diversité des formes linguistiques, par exemple le problème de la polysémie, nous avons choisi d'effectuer le travail de reconnaissance d'occurences de concepts au sein des textes. La présence d'un concept est reconnue par la présence d'une forme syntaxique particulière, ou bien d'un terme particulier (Kodratoff 2001; Fontaine et Kodratoff 2003).

Ainsi, dans la section 2.4, nous expliquerons notre méthodologie d'extraction de la terminologie du domaine. Les relations syntaxiques qui sont également considérées comme des instances des concepts sont traitées dans (Fontaine et Kodratoff 2003).

L'utilisation des ontologies construites dans la première phase du processus de fouille de textes permet une représentation condensée et simplifiée des corpus que nous étudions. L'utilisation de telles ontologies permet de représenter le corpus selon une matrice numérique *texte* $\times$ *concept*. La seconde phase de notre étude consistera