Exploitation de l'asymétrie entre termes pour l'extraction automatique de taxonomies à partir de textes

Davide Buscaldi*, Guillaume Cleuziou*
Gaël Dias** Vincent Levorato*,***

*LIFO, Université d'Orléans
{davide.buscaldi,guillaume.cleuziou,vincent.levorato}@univ-orleans.fr
**GREYC, Université de Caen Basse-Normandie
gael.dias@unicaen.fr
***IRISE, CESI
vlevorato@cesi.fr

Résumé. Nous présentons dans cet article une nouvelle approche pour la génération automatique de structures lexicales (ou taxonomies) à partir de textes. Cette tâche est fondée sur l'hypothèse forte selon laquelle l'accumulation de faits statistiques simples sur les usages en corpus permet d'approximer des informations de niveau sémantique sur le lexique. Nous utilisons la prétopologie comme cadre de travail afin de formaliser et de combiner plusieurs hypothèses sur les usages terminologiques et enfin de structurer le lexique sous la forme d'une taxonomie. Nous considérons également le problème de l'évaluation des taxonomies résultantes et proposons un nouvel indice afin de les comparer et de positionner notre approche par rapport à la littérature.

1 Introduction

Le codage des relations sémantiques entre concepts au sein d'une structure lexico-sémantique de type « taxonomie » permet d'améliorer considérablement la pertinence des processus de recherche d'information (RI) et de traitement automatique des langues (TAL). Cependant, l'utilisation de telles ressources est fortement limitée du fait des efforts considérables à entreprendre pour les construire. Afin de limiter ces efforts, un certains nombre de recherches ont été entreprises ces dernières années pour « apprendre » des taxonomies à partir de l'observation des usages en corpus (Biemann, 2005; Cimiano et al., 2009). L'apprentissage automatique de taxonomies à partir de textes plutôt que leur construction manuelle présente des avantages indéniables. Non seulement cela permet d'ajuster la connaissance extraite à tout domaine de spécialité en choisissant le corpus adapté ; mais de plus, le coût par entrée lexicale sera considérablement réduit par rapport à une décision experte manuelle ou même assistée, permettant ainsi la génération de ressources plus importantes.
Différentes méthodologies d'apprentissage ont été proposées afin de construire automatiquement des taxonomies. Elles peuvent être organisées en trois types d'approches : les méthodes fondées sur la notion de similarités (Paaß et al., 2004; Cimiano et al., 2004), sur la théorie des