

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.

Summary

We present in this paper a new approach for the automatic generation of lexical structures from texts. This task is based on the strong hypothesis that simple statistical observations on textual usages can provide pieces of semantics about the lexicon. Using such "naive" observations only, we propose a (pre)-topological framework to formalize and combine various hypotheses on textual data usages and then to derive a structure similar to usual lexical knowledge databases such as WordNet. In addition, we also consider the evaluation problem for the resulting lexical structures; we propose a new measure to compare two structures and use it to quantify the contribution of the new structuring approach with respect to the corresponding solution proposed by ? on two case studies that differ on the domain and the size of the lexicon.