Traitement automatique d'informations textuelles complexes : connaissances linguistiques hétérogènes et à granularité variable

Marion Laignelet

Université Toulouse 2, Laboratoire CLLE-ERSS
5 allée Antonio Machado, 31058 Toulouse Cedex 9
marion.laignelet@univ-tlse2.fr

Résumé. Dans cet article, nous présentons une méthodologie permettant le traitement et la structuration de données linguistiques complexes. Par données complexes, nous envisageons des informations textuelles présentant la particularité d'être à la fois hétérogènes sémantiquement et à granularité variable. Pour passer d'une structure linguistique constituée d'objets complexes à une organisation des données permettant l'application de méthodes statistiques et/ou de fouille de données, nous proposons un modèle de représentation des unités du discours. Ce travail est mené dans le cadre d'un projet visant la mise en oeuvre d'un prototype d'aide à la mise à jour de documents encyclopédiques articulé autour du repérage automatique de zones textuelles contenant de l'information obsolète.

1 Introduction

Dans cet article, nous proposons un modèle de représentation des unités discursives qui nous permet de passer d'annotations linguistiques complexes à une représentation des informations manipulables par des outils statistiques et/ou de fouille de données.

En linguistique, les objets étudiés sont divers et relativement stables au sein des communautés : le morphème pour les morphologues, le mot, lexie ou lexème pour le lexicologue, les catégories syntaxiques pour les syntacticiens (Brunet, 2003; Kao et Poteet, 2007). L'utilisation des statistiques en linguistique n'est pas nouvelle : les morphologues utilisent les statistiques pour prédire les schémas de formation des mots (Grabar et Zweigenbaum, 1999; Daille et al., 2001; Jacquemin, 1997; Lavallée et Langlais, 2010), les lexicologues et sémanticiens étudient les fréquences d'apparition de termes en collocation (Bouillon et al., 2000; Hearst, 1992; Sébillot et al., 2000) ou la recherche de motifs (Nouvel et al., 2010), les syntacticiens quant à eux s'intéressent aux structures récurrentes à l'intérieur des phrases (Brill, 1992; Schmid, 1994; Dejean et al., 2010).

D'un point de vue applicatif, ce travail a été élaboré dans le cadre d'un projet de recherche visant le repérage automatique de segments textuels contenant de l'information potentiellement obsolète. Repérer de tels segments permet *in fine* d'aider des rédacteurs à mettre à jour les contenus d'encyclopédies.