Extraction de termes centrée autour de l'expert

Thomas Heitz, Mathieu Roche, Yves Kodratoff

Université Paris-Sud, Bât 490, 91405 Orsay Cedex France,
{heitz, roche, yk}@lri.fr,
http://www.lri.fr/∼{heitz, roche, yk}/

Résumé. Nous développons un logiciel, Exit, capable d'aider un expert à extraire des termes qu'il trouve pertinents dans des textes de spécialité. Tout est mis en place pour faciliter le travail de l'expert afin qu'il puisse consacrer son temps à la seule reconnaissance des termes pertinents. Pour cela, différentes mesures statistiques et de nombreuses options d'extraction sont disponibles dans Exit. Afin d'utiliser au mieux les connaissances de l'expert, notre approche est semi-automatique. De plus, l'expert construit des termes pouvant inclure des termes précédemment extraits ce qui rend itératif et constructif notre processus de formation des termes. Enfin, l'ergonomie du logiciel a profité des enseignements tirés lors de son utilisation pour une compétition internationale d'extraction de connaissances.

1 Introduction

Le logiciel Exit (Roche et al., 2004b), **EX**traction **I**térative de la **T**erminologie permet l'extraction des collocations. Une collocation peut être définie comme *une combinaison de mots dont le sens global est déductible des unités qui la composent, une des unités caractérisant l'autre*. Par exemple : *plante à fleurs*, *à fleurs* caractérisant *plante*.

Toutes les expressions extraites par ce logiciel sont des collocations. Mais ce qui intéresse l'expert, ce sont les **collocations pertinentes** qui sont des expressions ayant un sens unique pour un domaine précis. Nous les appellerons **termes** dans la suite de cet article.

Exit est destiné à des utilisateurs experts d'un domaine. Ceux-ci doivent donc avoir une connaissance approfondie des notions employées dans les textes de spécialité analysés pour pouvoir reconnaître les collocations pertinentes parmi celles extraites par le logiciel.

Ce logiciel est un des modules d'une chaîne de fouille de textes (Mathiak et Eckstein, 2004; Roche et al., 2004a) qui comprend les étapes de normalisation, étiquetage grammatical, construction de la terminologie avec Exit, classification conceptuelle, etc. Ceci permet ensuite de pouvoir traduire, résumer, générer ou interroger des textes. Les entrées d'Exit correspondent à des textes étiquetés grammaticalement, notamment (Brill, 1995), et les sorties correspondent à des listes de termes qui peuvent être associés à des concepts grâce à des systèmes de construction de classifications conceptuelles (Kodratoff, 2004).

Un des points forts d'Exit est son processus itératif qui permet de construire des termes incluant des termes précédemment trouvés. De plus, tout le processus d'extraction est centré autour de l'expert et nous avons fait en sorte qu'il soit aidé au maximum pour qu'il passe le moins de temps possible à cette tâche.