Logiciel d'aide à l'étiquetage morpho-syntaxique de textes de spécialité

Ahmed Amrani*, Jérôme Azé**, Yves Kodratoff**
*ESIEA Recherche, 9 rue Vésale, 75005 Paris, France
amrani@esiea.fr
** LRI, Université Paris Sud, 91405 Orsay Cedex, France
{aze,yk}@lri.fr, http://www.lri.fr/~{aze,yk}

Résumé. La compréhension de textes de spécialité nécessite un étiquetage morpho-syntaxique de bonne qualité. Or, lorsque les textes étudiés sont issus de domaines spécifiques et peu usités, il est rare de disposer de dictionnaires et autres ressources lexicales fiables. Le logiciel que nous proposons permet d'utiliser un étiquetage réalisé par un étiqueteur généraliste, puis d'améliorer cet étiquetage en intégrant des connaissances d'experts du domaine étudié. Grâce au logiciel développé, il est relativement aisé pour un expert du domaine de détecter des erreurs d'étiquetage et de mettre en place des règles de ré-étiquetage. Ces règles peuvent être obtenues de deux manières différentes : (1) soit en utilisant un langage de programmation permettant d'exprimer des règles complexes de ré-étiquetage, (2) soit par apprentissage automatique des règles à partir d'exemples corrigés au moyen d'une interface dédiée. Cet apprentissage propose de nouvelles règles à l'expert, acquises automatiquement.

1 Introduction

La compréhension de textes de spécialité repose sur un étiquetage morpho-syntaxique de bonne qualité. Or, lorsque les textes étudiés sont issus de domaines spécifiques et peu usités, il est rare de disposer de dictionnaires et autres ressources lexicales fiables. Ainsi, les systèmes d'étiquetage (Brill, 1994; Schmid, 1994) ne sont pas en mesure d'étiqueter correctement des textes spécialisés. Ayant réalisé ce constat et face au besoin d'avoir des textes correctement étiquetés pour pouvoir en extraire des connaissances utiles, il devient indispensable de corriger l'étiquetage. De nombreux outils peuvent être utilisés pour modifier et corriger l'étiquetage d'un texte.

L'étiqueteur de Brill (Brill, 1994) offre la possibilité d'écrire des règles contextuelles qui seront utilisées pour modifier l'étiquetage réalisé par défaut. Cependant, les règles ainsi exprimées ne sont pas utilisables en dehors de l'étiqueteur de Brill.

INTEX [1], bien que non conçu pour cette tâche, pourrait être utilisé pour détecter des erreurs d'étiquetage et pour les corriger. L'utilisation d'INTEX implique de disposer de dictionnaires dédiés au domaine pour obtenir un premier étiquetage relativement correct. Or, comme nous l'avons précédemment évoqué, il est difficile d'obtenir de telles ressources.

Des outils d'analyse syntaxique profonde des textes, tels qu'INTEX, sont certes nettement plus fiables qu'une simple analyse syntaxique de surface. Par contre, le temps de

1. http://www.nyu.edu/pages/linguistics/intex/