Combinaison de l'approche inductive (progressive) et linguistique pour l'étiquetage morphosyntaxique des corpus de spécialité

Ahmed Amrani*,**, Yves Kodratoff**
*ESIEA Recherche, Pôle ECD, 11 rue Baudin, 74200 Ivry sur Seine, France
amrani@esiea.fr
**LRI, UMR CNRS 8623, Bât. 490, Université de Paris-Sud 11, 91405 Orsay, France
yk@lri.fr

Résumé. Les étiqueteurs morphosyntaxiques sont de plus en plus performants et cependant, un véritable problème apparaît lorsque nous voulons étiqueter des corpus de spécialité pour lesquels nous n'avons pas de corpus annotés. La correction des ambiguïtés difficiles est une étape importante pour obtenir un corpus de spécialité parfaitement étiqueté. Pour corriger ces ambiguïtés et diminuer le nombre de fautes, nous utilisons une approche itérative appelée *Induction Progressive*. Cette approche est une combinaison d'apprentissage automatique, de règles rédigées par l'expert et de corrections manuelles qui se combinent itérativement afin d'obtenir une amélioration de l'étiquetage tout en restreignant les actions de l'expert à la résolution de problèmes de plus en plus délicats. L'approche proposée nous a permis d'obtenir un corpus de biologie moléculaire « correctement » étiqueté. En utilisant ce corpus, nous avons effectué une étude comparative de quatre étiqueteurs supervisés.

1 Introduction

Dans le cadre d'un processus complet de fouille de textes (Kodratoff et al., 2003, Amrani et al., 2004a), nous nous sommes intéressés à l'étiquetage morphosyntaxique des corpus de spécialité. L'étiquetage morphosyntaxique consiste à affecter à chaque mot dans la phrase son étiquette morphosyntaxique, en prenant en considération le contexte et la morphologie de ce mot. L'étiquette morphosyntaxique est composée de la catégorie syntaxique du mot (nom commun, nom propre, adjectif, etc.) et souvent comporte des informations morphologiques (genre, nombre, personne, etc.). Les outils informatiques nécessaires à l'opération d'étiquetage sont appelés « étiqueteurs ».

Un problème se pose lorsque les étiquettes des mots sont ambiguës. Par exemple, le mot *functions* peut être un nom au pluriel ('*biological functions are…*') ou bien un verbe au singulier ('*this gene functions as…*'). Le problème à résoudre est celui de trouver l'étiquette correcte selon le contexte. La correction de ces ambiguïtés est une étape importante pour obtenir un corpus de spécialité « parfaitement » étiqueté. Pour lever ces ambiguïtés et donc diminuer le nombre de fautes d'étiquetage, nous proposons une approche interactive et itérative appelée *Induction Progressive*. Cette approche est une combinaison d'apprentissage automatique, de règles rédigées par l'expert et de corrections manuelles. L'induction pro-