ETIQ, un étiqueteur inductif convivial pour les corpus de spécialité

Ahmed Amrani *, Oriane Matte-Tailliez**, Yves Kodratoff**
*ESIEA Recherche, 9 rue Vésale, 75005 Paris
ahmed.amrani@esiea.fr
**LRI, Bât. 490, Université de Paris-Sud 11, 91405 Orsay
oriane.matte@lri.fr, yves.kodratoff@lri.fr

Les travaux que nous présentons se rapportent à l'étiquetage grammatical de textes de spécialité. Les étiqueteurs existants sont entraînés sur divers corpus, et produisent des erreurs dans l'étiquetage des textes de spécialité, notamment les textes techniques et scientifiques. La méthode la plus triviale pour apprendre des automates adaptés à un domaine spécialisé est d'étiqueter manuellement de grands corpus du domaine, ce qui nécessite un travail fastidieux et un coût important. Pour éviter cela, nous proposons une méthode semi-automatique. ETIQ, le nouvel étiqueteur que nous avons construit, permet de corriger la base de règles obtenue par l'étiqueteur de BRILL [Brill, 1994] et de l'adapter à un corpus de spécialité. L'expert du domaine visualise l'étiquetage de base et le corrige par l'insertion de règles lexicales et contextuelles expressives et spécialisées.

Dans le module lexical, l'objectif est de trouver des règles lexicales spécialisées pour déterminer l'étiquette la plus probable pour tout mot de la spécialité.

Dans le module contextuel, des règles contextuelles peuvent être utilisées pour améliorer l'étiquetage. Ces règles corrigent l'étiquette du mot en fonction de son contexte, c'est-à-dire le mot lui-même, son étiquette, les mots voisins et leurs étiquettes.

Que ce soit dans le module lexical ou contextuel, le système permet à l'expert de visualiser les mots dans leur contexte, d'insérer une règle, de vérifier le résultat de son application et éventuellement de la modifier. Les règles utilisées sont plus modulables que celles de Brill. La grammaire de nos règles permet de combiner, par des opérateurs logiques, les conditions simples utilisées par Brill ou encore d'utiliser des expressions régulières.

Afin d'assister l'expert efficacement dans la tâche d'écriture de règles, nous proposons une approche inductive. Le programme d'induction utilisé (C4.5) prend en considération les améliorations de l'expert pour lui proposer de nouvelles règles. Nous avons appliqué cette approche dans la phase lexicale où nous avons utilisé les descripteurs les plus pertinents du domaine. Les règles obtenues de façon automatique sont triées par ordre décroissant d'une mesure (qui comprend couverture et précision). L'expert introduit enfin les règles qu'il juge « les meilleures ».

En effet, en utilisant des techniques d'apprentissage et en permettant à l'expert d'incorporer les connaissances du domaine de manière interactive, nous améliorons nettement l'étiquetage des corpus de spécialité. Pour notre corpus de biologie moléculaire, les corrections ont porté sur 8,6% de l'ensemble des étiquettes avec un taux de réussite, évalué par l'expert à 98,6%.

Référence

[Brill, 1994] E. Brill Some Advances in Transformation-Based Part of Speech Tagging. AAAI, 1:722-727.