Expérimentation en apprentissage d'heuristiques pour l'analyse syntaxique

Sylvain DELISLE
Département de mathématiques et d'informatique
Université du Québec à Trois-Rivières
Trois-Rivières, Québec, Canada, G9A 5H7
Sylvain_Delisle@uqtr.uquebec.ca

Sylvain LÉTOURNEAU, Stan MATWIN
School of Information Technology and Engineering, University of Ottawa
Ottawa, Ontario, Canada, K1N 6N5
sletour@ai.iit.nrc.ca , stan@site.uottawa.ca

Les systèmes ou programmes de traitement de la langue naturelle doivent prendre des décisions quant au choix des meilleures stratégies ou règles à appliquer en cours de résolution d'un problème particulier. Pour un analyseur syntaxique constitué d'une base de règles symboliques, le cas auquel nous nous intéressons ici, ces décisions peuvent consister à sélectionner les règles ou l'ordonnancement de celles-ci permettant de produire la plus rapide ou la plus précise analyse syntaxique pour un énoncé, un type d'énoncé ou même un corpus spécifique. La complexité de telles bases de règles grammaticales et leurs subtilités computationnelles et linguistiques font en sorte que la prise de ces décisions constitue un problème difficile. Nous nous sommes donc fixé comme objectif de trouver des techniques qui permettraient d'apprendre des heuristiques performantes de prise de décision afin de les incorporer à un analyseur syntaxique existant. Pour atteindre une telle adaptabilité, nous avons adopté une approche d'apprentissage automatisé supportée par l'utilisation de systèmes de classification automatique.

Nos travaux ont été réalisés sur un analyseur syntaxique à large couverture syntaxique de l'anglais écrit et ont porté sur un sous-ensemble précis de celui-ci : le niveau le plus haut qui doit décider avec quelle(s) règle(s)—et, s'il y en a plusieurs, dans quel ordre—lancer l'analyse syntaxique de l'énoncé en cours de traitement, selon que cet énoncé semble comporter des phénomènes de coordination structurelle plus ou moins compliqués. Ce problème de décision se traduit naturellement en un problème de classification, d'où notre utilisation de systèmes de classification automatique de plusieurs types : règles de décision, basé sur les instances, réseaux de croyances et réseaux de neurones. Soulignons que notre analyseur syntaxique possédait déjà des règles heuristiques dédiées à ce problème de décision. Elles avaient été composées par le premier auteur sans avoir recours à aucun mécanisme automatique. Nous désirions maintenant trouver de nouvelles heuristiques qui seraient encore plus performantes que les anciennes et qui pourraient donc les remplacer.

La méthodologie que nous avons utilisée est la suivante. Premièrement, nous avons défini les attributs les plus pertinents pour représenter les exemples (énoncés). Il importait d'identifier des attributs facilement calculables de façon automatique et qui permettraient d'obtenir de nouvelles heuristiques intéressantes. Par exemple, la présence de conjonctions de coordination et la longueur de l'énoncé sont deux attributs utiles. Deuxièmement, nous avons soumis les exemples, traduits en termes des attributs sélectionnés, aux systèmes classificateurs afin d'obtenir des règles. Nous avons ensuite sélectionné les règles les plus intéressantes, c'est-à-dire celles qui étaient les plus discriminantes tout en demeurant intelligibles dans une perspective linguistique. Troisièmement, nous avons incorporé les règles sélectionnées à notre analyseur syntaxique en remplacement des anciennes. Finalement, nous avons évalué la nouvelle version de l'analyseur obtenue grâce à ces nouvelles règles et effectué une comparaison avec l'ancienne version. Les résultats que nous avons obtenus se résument ainsi : nous avons trouvé de nouvelles heuristiques qui sont significativement meilleures que les anciennes et qui, en particulier, possèdent un taux d'erreur de 35% inférieur à celui des anciennes. Qui plus est, ces résultats ont été obtenus sur des énoncés tout à fait indépendants de ceux utilisés pour l'entraînement avec les systèmes classificateurs. Ces résultats démontrent que des techniques d'apprentissage automatisé peuvent concourir à l'optimisation adaptive de certaines décisions importantes en analyse syntaxique.