Analyse syntaxique des langues naturelles par combinaison de grammaires algébriques et décomposition lagrangienne

Joseph Le Roux*[1], Antoine Rozenknop[1], and Jennifer Foster[2]

[1]Laboratoire d'Informatique de Paris-Nord (LIPN) – Université Paris XIII - Paris Nord, CNRS : UMR7030, Institut Galilée – Institut Galilée 99, avenue J.B Clément 93430 VILLETANEUSE, France
[2]School of Computing (DCU) – Dublin City University, Glasnevin, Dublin 9, Irlande

Résumé

La décomposition fait désormais partie de la /trousse à outils/ formelle en traitement automatique des langues, en particulier en analyse syntaxique (Rush et Collins, 2010). Elle permet en effet de pallier le problème majeur de la taille de l'espace de recherche causé par l'ambigu´ité massive du langage naturelle, que la programmation dynamique ne permet pas toujours de résoudre, tout en donnant des /certificats d'optimalité/ aux solutions retournées, contrairement aux approximations à base de seuil habituellement utilisées.

La plupart des travaux dans ce domaine se sont consacrés soit à l'analyse syntaxique en dépendances d'ordre supérieur soit à la modélisation de tâches jointes (analyse syntaxique et étiquetage en partie du discours par exemple), et (à notre connaissance) il n'existe pas de travaux sur la décomposition pour l'analyse en constituants.

Nous présentons un algorithme qui permet de calculer la meilleure analyse à partir de plusieurs grammaires pondérées qui peuvent engendrer des langages différents mais "proches" modulo certaines opérations simples (renommage des nœuds et débinarisation des règles). Notre méthode repose sur la superposition partielle des meilleures solutions de chaque analyseur. Nous utilisons un algorithme de décomposition à base de sous-gradient projeté inspiré de l'algorithme d'inférence dans les champs markoviens aléatoires de (Komodakis et al, 2007).

Nous montrons expérimentalement que cette méthode permet d'améliorer les performances d'un système d'analyse syntaxique déjà très "compétitif". Nous obtenons des résultats /état-de-l'art/ sur le Penn Treebank, corpus de référence en analyse syntaxique, avec un F-score

*Intervenant

supérieur à 92,4.

URL: http://aclweb.org/anthology//D/D13/D13-1116.pdf

Mots-Clés: Natural Language Processing, Decomposition