Apprentissage non supervisé de dépendances syntaxiques à partir de texte étiqueté, plusieurs variantes de PCFG légères

Marie Arcadias*, Guillaume Cleuziou**,***, Edmond Lassalle*, Christel Vrain**

*Orange Labs, 2 avenue Pierre Marzin, 22300 Lannion
prénom.nom@orange.com,
**LIFO, Université d'Orléans, Rue Léonard de Vinci, 45067 Orléans Cedex 2
prénom.nom@univ-orleans.fr
***GREYC, UCBN, Bd. Maréchal Juin, 14032 Caen Cedex 5

Résumé. L'apprentissage de dépendances est une tâche consistant à établir, à partir des phrases d'un texte, un modèle de construction d'arbres traduisant une hiérarchie syntaxique entre les mots. Nous proposons un modèle intermédiaire entre l'analyse syntaxique complète de la phrase et les sacs de mots. Il est basé sur une grammaire stochastique hors-contexte se traduisant par des relations de dépendance entre les catégories grammaticales d'une phrase. Les résultats expérimentaux obtenus sur des benchmarks attestés dépassent pour cinq langues sur dix les scores de l'algorithme de référence DMV, et pour la première fois des scores sont obtenus pour le français. La très grande simplicité de la grammaire permet un apprentissage très rapide, et une analyse presque instantanée.

1 Introduction et état de l'art

Une structure de dépendances d'une phrase traduit une hiérarchie syntaxique des mots, et permet d'en inférer une sémantique. Les applications liées aux structures de dépendance sont multiples, on peut citer entre autres la modélisation de langages, la reconnaissance d'implications textuelles, les moteurs de question/réponse, l'extraction d'information, l'induction d'ontologies lexicales et la traduction automatique.

Une structure de dépendances d'une phrase (cf. Fig. 1 à gauche) est un arbre dont les nœuds sont les mots, ou tokens, de la phrase. Un des mots est désigné comme la racine de l'arbre (en général un verbe), à laquelle sont attachés des sous-arbres couvrant des portions contigües de la phrase. Un arbre de dépendances est constitué de relations orientées entre un mot syntaxiquement plus fort (tête) et un mot plus faible (dépendant). Le modèle de dépendances est un compromis intéressant entre l'analyse syntaxique classique complète et une représentation "sac de mots".

Les modèles d'apprentissage supervisé de dépendances exigent un nombre important d'exemples annotés à la main. Ce travail, très long et fastidieux, demande une expertise de linguiste essentielle et doit être remanié profondément à chaque nouveau type de texte analysé. La quantité de textes annotés en dépendances est faible comparée à la variété des types de textes disponibles sur le Web. Nous proposons une approche non supervisée ne demandant