Acquisition automatique de grammaires de contraintes

Jean-Philippe Prost

5 novembre 2012

Résumé

L'objet de ce sujet est de développer une stratégie d'acquisition automatique d'une grammaire de contraintes à partir d'un corpus annoté.

En matière de traitement automatique des langues naturelles, le développement de grammaires dites *couvrantes* (i.e. couvrant un large spectre de phéonèmes linguistiques d'une langue) est une tâche généralement effectuée manuellement. C'est une tâche extrêmement fastidieuse, et couteuse. Pour ces raisons (entre autres), c'est une ressource très rare.

Parallèlement, les corpus annotés avec des structures syntaxiques sont, eux, beaucoup plus nombreux et accessibles. Ces annotations ont, généralement, été produites manuellement, ce qui, dans un contexte linguistique, constitue une référence en termes qualitatifs. L'idée sous-jacente est de faire appel aux connaissances grammaticales de plusieurs annotateurs plutôt qu'à une grammaire pré-établie. Pour effectuer ces annotations, les annotateurs sont donc munis d'un *guide d'annotation*, qui se contente de fournir des directives générales de façon à assurer un minimum d'homogénéité dans les résultats inter-annotateurs.

Ces corpus sont très utiles aux traitement des langues par apprentissage automatique, ou encore pour l'évaluation qualitative de différents outils de TAL, en servant de référence.

Néanmoins, les règles de grammaire en application dans ces corpus restent implicites. Certains travaux de recherche ce sont donc attachés à les extraire automatiquement []. Par voie de conséquence, le résultat fournit une grammaire utilisable, plus ou moins directement, par différents outils de TAL qui en requièrent une. Cependant, les formalismes grammaticaux étant très nombreux, les mécanismes d'acquisition mis en oeuvre dans ces travaux, généralement pour un formalisme ou une famille de formalisme donnés, ne sont pas toujours aisément transposables dans un cadre formel différent que celui utilisé pour l'extraction.

L'objet de ce stage est donc de développer une stratégie d'acquisition applicable à une famille de formalismes particulière. Cette stratégie sera implémentée et testée sur corpus. L'évaluation s'effectuera par le biais de l'utilisation d'un analyseur syntaxique qui, muni de la grammaire acquise, devra reproduire l'annotation initiale.

Le stage s'attachera également à apporter des éléments de réponse (ou répondre) aux questions suivantes :

- Quelles sont les propriétés formelles de la grammaire extraite (par exemple, est-elle LL1) ?
- Les annotations sont-elles suffisantes pour prendre en compte tous les éléments d'information linguistique qui interviennent dans une grammaire couvrante (par exemple, règles d'accord) ?
- Dans l'hypothèse où les annotations seraient insuffisantes, est-il possible d'établir des règles d'acquisition générales basées sur une autre source d'information (par exemple, en utilisant un outil de pré-traitement morphologique, dans le cas des règles d'accord) ?