Un correcteur sémantique du français

Éric Violard (Eric.Violard@inria.fr)

Sujet de TER

Les correcteurs orthographiques sont des outils très utilisés et très pratiques lorsqu'il s'agit de rédiger des documents ou des courriels importants. Certains outils de traitement de texte contiennent même des correcteurs grammaticaux [2] capables d'analyser la structure des phrases et de repérer des fautes d'accord. Ces outils sont conçus sur des principes qui ont des ressemblances frappantes avec ceux utilisés dans les phases d'analyse lexicale et syntaxique d'un compilateur d'un langage de programmation.

Ce TER vise à étudier la faisabilité d'un correcteur de phrases en français non plus seulement orthographique ou grammatical, mais "sémantique", c'est-à-dire capable de détecter et lever des ambiguïtés dans une phrase en fonction du sens attaché aux mots : les phrases considérées étant par ailleurs orthographiquement et grammaticalement bien construites. Par exemple, la phrase "J'ai fait un seau dans le temps." est orthographiquement et grammaticalement correcte, mais sémantiquement incorrecte, car il y a une incohérence forte entre le mot "seau" et le complément "dans le temps" : un correcteur idéal pourrait éventuellement proposer "J'ai fait un saut dans le temps." ou bien suggérer à l'utilisateur d'ajouter des mots à la phrase pour préciser le sens. Par ailleurs, la phrase "Un monde parfait est sans aucune ambiguïté." est une autre phrase sans erreur sémantique a priori.

Il s'agira bien sûr d'effectuer une recherche bibliographique préalable sur les outils existants dans ce domaine : leurs principes sous-jacents, leurs capacités, leurs limitations. L'étudiant s'intéressera en particulier à des correcteurs basés sur les idées suivantes : – utilisation du web qui constitue en soit une énorme base de données pour établir un lien sémantique entre les mots d'une phrase ou pour établir la probabilité [1] qu'une certaine association de mots soit correcte en se basant sur le nombre d'occurrences ; – analogie avec la structure d'un compilateur et en particulier le découpage en trois phases d'analyse : analyse lexicale (= orthographe), analyse syntaxique (= grammaire), analyse sémantique (= signification). Dans ce découpage, la correction sémantique d'un texte en français correspond au contrôle de type d'un programme.

Références

[1] Philip S. Kernick and David M. W. Powers. A Statistical Grammar Checker, 1996.

[2] Markus Malmsten Lund. Grammar Checker, 2004.

Référence clé (pour l'UE Initiation à la recherche) : [1]