Méthode pour l'analyse automatique de structures formelles sur documents multilingues

Cette thèse traite de l'analyse automatique de structures formelles de l'écrit. Elle commence par une excursion dans le multilinguisme au cours de laquelle nous présentons les documents dans leur dimension multilingue et montrons la nécessité de les traiter comme tels. Nous étudions leur structure multilingue et développons comment la calculer à l'aide d'un identificateur de langues.

Nous poursuivons par l'exposé d'une méthode originale d'analyse syntaxique automatique d'énoncés français tout-venants. Cette méthode est issue de nos travaux de généralisation et d'abstraction des recherches de Jacques Vergne. Les structures syntaxiques auxquelles nous nous sommes particulièrement intéressé sont le syntagme minimal et la proposition ; deux unités auxquelles il est possible d'associer une définition ayant une validité multilingue, ce qui rend la méthode applicable à diverses langues.

Nous proposons deux processus permettant la construction de ces unités. Ces processus considèrent les énoncés comme des flux textuels et construisent chacun leurs structures syntaxiques par propagation de contraintes relationnelles. Les structures intra-syntagmatique et intra-propositionnelle étant dépendantes, elles sont construites par l'interaction des deux processus, le second processus acceptant de travailler sur des unités partiellement définies. Enfin, nous montrons que les deux processus sont identiques si l'on fait abstraction de la nature de l'unité qu'ils construisent et de la base de règles qu'ils manipulent.

Le fil conducteur de cette thèse est la méthode. À chaque calcul de structure, nous mettons en effet l'accent sur la méthode ayant permis son obtention. Nous montrons que cette méthode est unique. Chaque structure est en effet calculée à partir d'indices formels et positionnels à la fois internes et externes : internes par l'étude des unités qui composent la structure, externes par l'étude du rôle de cette structure dans l'unité qui l'intègre.

Mots-clés : langage naturel, traitement du (informatique)/analyse documentaire/analyse automatique (linguistique)/multilinguisme

Discipline : Informatique

GREYC CNRS UPRESA 6072
Groupe de Recherche en Informatique, Image, et Instrumentation de Caen
Université de Caen — Campus II — F14032 Caen Cedex