UNE EXPERIENCE PRATIQUE D'UTILISATION DE L'ANALYSE LINGUISTIQUE EN RECHERCHE D'INFORMATION : BILAN & PERSPECTIVES

Ernest GRANDJEAN, Gerard VEILLON

Laboratoire IMAG
BP 53X - 38041 GRENOBLE cédex
France

Résumé

I - PRINCIPE DE L'APPLICATION PIAFDOC

I. 1. Méthode d'analyse textuelle assistée par ordinateur

Le programme PIAF est constitué par un ensemble de modules d'analyse linguistique. Déjà présenté par ailleurs, nous rappellerons que l'objectif était d'obtenir un outil suffisamment performant pour permettre l'analyse du texte libre, en faisant appel à un principe d'interaction avec l'utilisateur. En particulier, il est toujours possible de modifier grammaires et dictionnaires en cours d'analyse.

I. 2. Problèmes posés par les systèmes documentaires

Les techniques d'indexation automatique fondées sur un 'antidictionnaire' conduisent à reconnaître pour mots-clés toutes les variantes linguistiques de la même unité lexicale et ne traite pas les locutions. De plus, la moindre erreur typographique peut conduire à un mot-clé erroné. A l'interrogation, il n'est pas possible de tenir compte des fonctions syntaxiques ou des relations entre mots.

I. 3. Fonctions du module linguistique pour l'indexation et l'interrogation

Le programme PIAFDOC, dérivé de PIAF, a pour rôle de controler la conformité des données textuelles, de choisir pour chaque unité lexicale un représentant, qui peut d'ailleurs être le représentant d'une classe de synonymes, et de traiter une partie des groupes de mots ou locutions. Pour cela, il doit posséder un lexique complet du vocabulaire de l'application. Ce programme est implanté et exploité sur un centre serveur et disponible sur le réseau TRANSPAC. Il est expérimenté pour la constitution d'une base de données politiques. A l'interrogation, le même procédé doit conduire à réutiliser le même ensemble de mots-clés par un traitement de la question identique à celui du texte.

II - ANALYSE DES RESULTATS D'UNE APPLICATION REELLE

II. 1. Controle et validité des données textuelles

L'analyse sémantique du texte est limitée par les ambiguités inhérentes à tout système formel. Le recours à l'utilisateur ne devrait intervenir qu'en cas de réelle polysémie, ou d'insuffisance du lexique. La fréquence des interactions pourrait ainsi être réduite.

II. 2. Choix des descripteurs et représentation de l'information

Il est difficile de définir exactement les critères de choix des mots-clés. Une tendance naturelle à préciser le plus possible le contenu du texte peut conduire à tenir compte de constructions linguistiques complexes inaccessibles dans l'analyse du texte libre, en contradiction avec le principe d'indexation automatique.

II. 3. Rôle de l'interrogation

L'interrogation doit faire appel au même traitement que l'indexation, afin de faire référénce à un ensemble de mots-clés normalisés communs. Cependant, l'indexation systématique peut devenir bruyante, ou fournir des unités documentaires trop longues. Il faut alors une relecture du texte à l'interrogation pour ne retenir que les unités documentaires valides.

III - PROPOSITIONS POUR UN SYSTEME D'INFORMATIONS TEXTUELLES

III. 1. Fonctions d'un programme d'indexation et d'interrogation

III. 1. 1. Pour éviter toute ambiguité dans le choix des mots, la seule solution consiste à prendre une indexation minimale, indépendante du domaine considéré, complétée par une analyse morphosyntaxique conversationnelle qui doit permettre de déterminer les parties du discours dans la majorité des cas.

III. 1. 2. Interprétation des questions et des textes sélectionnés.

L'interrogation en langue naturelle conduit tout d'abord à une indexation identique à celle de la création. L'ensemble des textes ainsi retenus sont ensuite parcourus par des techniques algorithmiques efficaces pour isoler les mots ou groupes de mots ayant servi à les sélectionner. Une analyse linguistique plus fine doit permettre de vérifier la cohérence entre le segment de texte et la question.

III. 2. Implantation répartie : décentralisation des fonctions de création et d'accès par rapport aux bases

Le module linguistique peut raisonnablement être implanté sur petit matériel, pour permettre ainsi un traitement local des textes ou des questions indépendant des bases de données utilisées.