ANALYSEUR DE CORPUS POUR LANGUES FAIBLEMENT FLEXIONNELLES (EX: JAPONAIS, CHINOIS)

APPLICATION A LA RECHERCHE D'INFORMATIONS

Raoul Blin (CRLAO – CNRS)
blin @ ehess . fr

Nous présentons le logiciel SAGACE, analyseur de corpus pour les langues faiblement flexionnelles (par exemple japonais ou chinois).

Ses fonctionnalités permettent de l'utiliser pour l'extraction d'informations à partir de grands corpus (tout document texte numérisé, et en particulier pages web), en exploitant des relations sémantiques grammaticalisées.

Conceptions et fonctionnalités de base du logiciel

Le logiciel est constitué d'un moteur de recherche sur corpus associé à un lexique. Une des particularités du moteur est de travailler sur des corpus non tagués, et d'être indépendant du codage des textes (possibilité donc de travailler avec les encodages les plus courants pour les textes asiatiques).

1) extraction / comptage de collocations
2) extraction de collocations et de leur contexte

En vue d'applications

- mise en oeuvre immédiate sur texte brute (pas de pré-traitement nécessaire)
- disponibilité des dictionnaires (lexiques disponibles pour le japonais, 250.000 entrées, et le chinois, 70.000 entrées)
- très grande configurabilité

Exemple d'applications (autres que linguistiques) : recherche de noms de médicaments et de leur fabricant dans les textes japonais

Repose sur les observations linguistiques suivantes: la relation sémantique "le médicament M fabriqué par l'entreprise E" est grammaticalisée sous la forme d'un groupe nominal de la forme "S の E" ("*le* E *de* S").

La recherche est facilement organisée en fonction des données disponibles :

1) Recherche avec informations complètes : connaissant les noms d'entreprises (E) et les noms du médicament (M), il s'agit d'une simple recherche de collocations; SAGACE renvoie la liste de tous les couples <S-E> trouvés dans la structure S の E.

2) Recherche avec connaissances lacunaires sur un élément ; par exemple, si les noms d'entreprises ne sont pas disponibles, ils peuvent être repérés grâce à une analyse morphologique simple : les noms d'entreprises sont souvent constitués d'un nom propre

"<nom de ville, de personne> + <suffixe 社/株式 ("*Corp.*")>

Disponibilité du logiciel et des lexiques

Le logiciel SAGACE est en libre distribution et open source. Il est en usage depuis 2 ans pour les recherches en japonais et depuis un an pour le chinois. Les lexiques LEXS sont en libre distribution.