Extraction et filtrage de syntagmes nominaux pour la Recherche d'Information

Chedi Bechikh Ali*, Hatem Haddad*,**

*Faculté des Sciences de Tunis
Département Infromatique
Laboratoire LIPAH
Campus Universitaire El Manar, Tunis
Tunisie
chedi.bechikh@gmail.com,
**ESSTHS, Université de Sousse
H. Sousse, Tunisie
haddad.hatem@gmail.com

Résumé. Nous proposons dans cet article un Système de Recherche d'Information (SRI) qui se base sur des techniques d'indexation de textes en langue naturelle. Nous présentons une méthode d'indexation de documents qui repose sur une approche hybride pour la sélection de descripteurs textuels. Cette approche emploie des traitements du langage naturel pour l'extraction des syntagmes nominaux et sur un filtrage statistique basé sur l'information mutuelle pour sélectionner les syntagmes nominaux les plus informatifs pour le processus d'indexation. Nous effectuons des expérimentations en utilisant le corpus Le Monde 94 de la collection CLEF 2001 et sur le SRI Lemur pour évaluer l'approche proposée.

1 Introduction

La plupart des Systèmes de Recherche d'Information (SRI) utilisent des termes simples pour indexer et retrouver des documents. Cependant, cette représentation n'est pas assez précise pour représenter le contenu des documents et des requêtes, du fait de l'ambigüité des termes isolés de leur contexte : si l'on considère le mot composé « *pomme de terre* », les mots simples pomme et terre ne gardent pas leur propre sens que dans l'expression « *pomme de terre* » et si on les utilise séparément ils deviennent une source d'ambigüité.

Une solution à ce problème consiste à utiliser des termes complexes à la place des termes simples isolés (Boulaknadel, 2006). L'hypothèse est que les termes complexes sont plus aptes à désigner des entités sémantiques que les mots simples et constituent alors une meilleure représentation du contenu sémantique des documents (Mitra et al., 1997).

Notre objectif consiste à acquérir des termes complexes représentatifs du contenu informationnel du corpus. Les termes complexes extraits doivent représenter le contenu des textes sous une forme compréhensive par l'ordinateur et riche en information. Ces termes extraits