Modèle de Langue à base de Concepts pour la Recherche d'Information

Lynda SAID L'HADJ*, Mohand BOUGHANEM**

*Ecole Doctorale STIC, Ecole nationale Supérieure d'Informatique ESI, Algérie
l_said_lhadj@esi.dz
**Laboratoire IRIT, Université Paul Sabatier
118 route de Narbonne 31062 Toulouse Cedex 09, France
bougha@irit.fr

Résumé. La majorité des modèles de langue appliqués à la recherche d'information repose sur l'hypothèse d'indépendance des mots. Plus précisément, ces modèles sont estimés à partir des mots simples apparaissant dans les documents sans considérer les éventuelles relations sémantiques et conceptuelles. Pour pallier ce problème, deux grandes approches ont été explorées : la première intègre des dépendances d'ordre surfacique entre les mots, et la seconde repose sur l'utilisation des ressources sémantiques pour capturer les dépendances entre les mots. Le modèle de langue que nous présentons dans cet article s'inscrit dans la seconde approche. Nous proposons d'intégrer les dépendances entre les mots en représentant les documents et les requêtes par les concepts.

1 Introduction

Les modèles de langue ont acquis une grande popularité en Recherche d'Information (RI) étant donné la solidité de leur fondement mathématique Ponte et Croft (1998). Ces modèles ne modélisent pas directement la notion de pertinence. Cette dernière est vue comme la probabilité conditionnelle ($P(Q|D)$) que la requête Q soit générée par le modèle de langue du document (D). $P(Q|D)$ est estimée sous l'hypothèse d'*indépendance* des mots qui simplifie le calcul mathématique. Cependant, elle pose un problème majeur lié à la représentation des documents (requêtes) comme des sacs de mots dénués de sémantique.

Pour pallier ce problème, une nouvelle génération de modèles de langue qui s'inscrit à l'intersection des modèles de langue et de la recherche sémantique d'information a été développée Cao et al. (2005), Srikanth et Srihari (2002, 2003). Dans cette intersection, deux grandes approches peuvent être distinguées : *l'approche statistique surfacique* qui prend en compte des dépendances surfaciques entre les mots et *l'approche sémantique* basée sur des ressources sémantiques (ontologies, thésaurus) pour identifier les sens des mots.

Le modèle de langue présenté dans cet article s'inscrit dans la seconde approche. Nous proposons de capturer les dépendances entre les mots par l'identification des concepts auxquels renvoient ces mots. Même si l'approche conceptuelle souffre du problème de silence[1], nous pensons que la définition d'un modèle de langue mixte combinant les concepts identifiés dans l'ontologie et les concepts non identifiés permet de résoudre ce problème.

[1] dû à la non disponibilité de ressources conceptuelles complètes et générales.