Application des vecteurs sémantiques à la fouille de texte

Jacques Chauché

LIRMM-CNRS et Université Montpellier 2
161 rue Ada, 34395 Montpellier cedex 5
chauche@lirmm.fr
http://www.lirmm.fr/ chauche

Résumé. L'approche présentée ici se base sur un traitement du contenu syntaxico-sémantique par un analyseur du Français, le système SYGFRAN(SYGFRAN) , pour retrouver un ensemble de phrases appartenant à différents discours du président François Mitterrand plongées dans un ensemble de phrases appartenant à différents discours du président Jacques Chirac. Ce traitement se fait par calcul de vecteurs sémantiques de phrases (méthodologie définie dans l'article) et par la définition d'une relation de similitude décrivant l'inclination de vecteurs dont l'inclinaison, ou distance angulaire, est proche. A l'aide de cette relation, des phrases sont attribuées par le système à l'un ou l'autre des auteurs, et l'article indique des F-mesures obtenues sur le premier corpus, dit d'apprentissage, légèrement supérieures à 80%.

1 Présentation

Le défi DEFT 2005 organisé pour le congrès annuel TALN consiste à retrouver un ensemble de phrases appartenant à différents discours du président François Mitterrand plongées dans un ensemble de phrases appartenant à différents discours du président Jacques Chirac. Les phrases introduites traitent d'une thématique distincte de la thématique retenue pour les phrases des discours de Jacques Chirac. Il est donc possible d'aborder ce problème par la recherche d'une distinction sémantique. L'approche présentée ici se fonde sur un traitement du contenu par opposition aux traitements habituels basés sur des approches statistiques. Le vocabulaire utilisé par l'un ou l'autre n'aura d'importance qu'à travers les idées qu'il véhicule. L'originalité de cette approche vient du fait qu'elle s'appuie sur la structure syntaxique du texte. L'analyse syntaxique produit cette structure syntaxique pour les groupes, les phrases et le texte. L'obtention de cette structure est très difficile et le taux de construction complète se situe autour de 30 %. Pour qu'une analyse soit complète il est nécessaire que la ou les structures (en cas d'ambiguïté) aient une seule racine. Dans les autres cas l'analyse produit une structure partielle qui permet néanmoins le calcul d'un vecteur sémantique avec une légère erreur. Dans ce cas le vecteur sémantique est obtenu à partir des groupes fonctionnels. Le taux de reconnaissance des groupes fonctionnels et des fonctions syntaxiques se situe autour de 70 % d'après les résultats de l'évaluation EASY. L'exposé de cette approche comprend trois parties : la méthode d'analyse syntaxique, la construction d'un vecteur sémantique associée à une structure et enfin l'utilisation de la suite des vecteurs pour identifier les passages rechercher.