Un duel probabiliste pour départager deux Présidents

Marc El-Bèze*, Juan-Manuel Torres-Moreno*,** Frédéric Béchet*

*Laboratoire d'Informatique d'Avignon - UAPV,
BP 1228, 84911 Avignon cedex 09, France
{ marc.elbeze, juan-manuel.torres, frederic.bechet }@univ-avignon.fr
http://www.lia.univ-avignon.fr
**École Polytechnique de Montréal - Département de Génie Informatique,
CP 6079 Succ. Centre-ville H3C 3A7 Montréal (Québec), Canada

Résumé. Nous présentons une palette de modèles probabilistes que nous avons employés dans le cadre du défi DEFT'05. La tâche proposée conjuguait deux problématiques distinctes du Traitement Automatique du Langage : l'identification de l'auteur (au sein de discours de Jacques Chirac, a pu être insérée une séquence de phrases de François Mitterrand) et la détection de ruptures thématiques (les thèmes abordés par les deux auteurs sont censés être différents). Pour identifier la paternité de ces séquences, nous avons utilisé des chaînes de Markov, des modèles bayésiens, et des procédures d'adaptation de ces modèles. Pour ce qui est des ruptures thématiques, nous avons appliqué une méthode probabiliste modélisant la cohérence interne des discours. Son ajout améliore les performances. Une comparaison avec diverses approches montre la supériorité d'une stratégie combinant apprentissage, cohérence et adaptation. Les résultats que nous obtenons, en termes de précision (0,890), rappel (0,955) et Fscore (0,925) sur le sous-corpus de test sont très encourageants.

1 Introduction

Dans le cadre des conférences TALN[1] et RECITAL[2] tenues en juin 2005 à Dourdan (France), un atelier a été organisé autour du défi de fouille textuelle proposé par (Azé et Roche, 2005). Ce défi portait le nom de DEFT'05 (DÉfi Fouille de Textes). Il a été motivé par le besoin de mettre en place des techniques de fouille de textes permettant soit d'identifier des phrases non pertinentes dans des textes, soit d'identifier des phrases particulièrement singulières dans des textes apparemment sans réel intérêt. Concrètement, il s'agissait de supprimer les phrases non pertinentes dans un corpus de discours politiques en français. Cette tâche est proche de la piste *Novelty* du challenge TREC (Soboro, 2004) qui dans sa première partie consiste à identifier les phrases pertinentes puis, parmi celles-ci, les phrases nouvelles d'un corpus d'articles journalistiques. Pour mieux comprendre à quoi correspondait dans DEFT'05 la suppression des phrases non pertinentes d'un corpus de discours politiques (Alphonse et al.,

[1]Traitement Automatique des Langues Naturelles.
[2]Rencontre des Etudiants Chercheurs en Informatique pour le Traitement Automatique des Langues.