Identification de thème et reconnaissance du style d'un auteur pour une tâche de filtrage de textes

Michèle Jardino*, Martine Hurault-Plantet*, Gabriel Illouz*

*LIMSI-CNRS, BP 133, 91403 ORSAY Cedex
{Michele.Jardino, Martine.Hurault-Plantet, Gabriel.Illouz}@limsi.fr
http://www.limsi.fr/Scientifique/lir

Résumé. Pour résoudre une tâche de filtrage des textes d'un auteur, insérés dans les textes d'un autre auteur, nous avons utilisé à la fois le style de l'auteur et la structure thématique du texte. Nous caractérisons le style d'un auteur par un modèle de langage n-grammes de mots ou de caractères entraîné sur un corpus d'apprentissage. Nous appliquons ensuite les modèles sur chaque phrase du corpus de test pour en calculer l'auteur le plus probable. Un algorithme de lissage transforme ensuite les résultats en segments continus pour chaque auteur. Parallèlement, nous avons élaboré une méthode d'identification du thème de chaque auteur dans un document. Nous déterminons d'abord les segments de texte de plus grande densité, pour chaque mot du document, par chaînage lexical. Puis, nous recherchons les chaînes lexicales principales des deux thèmes, par hypothèse celles dont les segments respectifs sont les plus étendus et se recouvrent le moins. Les résultats des deux méthodes sont finalement fusionnés.

1 Introduction

La tâche soumise à évaluation dans l'atelier DEFT'05 de la conférence TALN 2005 consistait à séparer les allocutions respectives de deux hommes politiques dans le même document. L'atelier d'évaluation s'est déroulé en deux temps : une phase d'entraînement pendant laquelle nous avons disposé d'un corpus d'apprentissage, puis une phase de test avec un nouveau corpus sur lequel l'évaluation proprement dite a été faite. Chaque corpus est composé d'allocutions de Jacques Chirac dans lesquelles des segments d'allocutions de François Mitterrand ont été glissés. Le nombre de phrases de Chirac est, dans chaque corpus, nettement plus important que le nombre de phrases de Mitterrand. L'évaluation de la tâche se fait par rapport à la reconnaissance des phrases de Mitterrand. La tâche ainsi définie est du filtrage de textes dans le sens où les phrases d'un auteur donné doivent être filtrées dans le flux des textes d'un autre auteur. Dans l'atelier d'évaluation, trois tâches étaient proposées pour trois versions différentes du corpus de test : une version avec des étiquettes à la place des noms propres et des dates, une version avec des étiquettes à la place des dates, et une version intégrale comportant les noms propres et les dates.

Deux critères peuvent aider à séparer les textes des deux auteurs dans chaque allocution : l'ensemble des caractéristiques d'écriture, propre à chaque auteur et qui représente son style, et les thèmes abordés. En effet, le thème de chaque allocution de Chirac a été choisi différent de celui du fragment d'allocution de Mitterrand qu'il contient. Nous avons donc essayé de