Modèles multi-thématiques markoviens pour la segmentation de textes

Loïs Rigouste*, Olivier Cappé*, François Yvon*
et Fabrice Clérot**

* GET – Télécom Paris & CNRS – LTCI
46 rue Barrault, 75634 Paris Cédex 13
rigouste,cappe,yvon à enst.fr

** France Télécom Division R & D TECH/SUSI/TSI
2 Avenue Pierre Marzin, 22307 Lannion Cédex
fabrice.clerot à francetelecom.com

Résumé. Dans cet article, nous montrons comment des outils génériques de la fouille statistique de textes peuvent être utilisés pour résoudre une tâche d'apprentissage supervisée: le DÉfi Fouille de Textes 2005. Dans un premier temps, nous étudions comment capturer une partie des spécificités de la tâche à l'aide de modèles de Markov cachés. Nous détaillons ensuite une modélisation des textes par un mélange de distributions multinomiales sur les comptes de mots, dans laquelle chaque composante correspond à un thème particulier. Les paramètres des distributions thématiques sont estimés grâce à l'algorithme EM. Ce modèle est utilisé pour diviser en sous-thèmes les discours des deux présidents. Nous discutons finalement des performances obtenues en combinant ces deux outils.

1 Introduction

La tâche DEFT, introduite plus en détail dans ce même numéro, consiste à analyser un pseudo-document construit en insérant, dans un discours de Jacques Chirac, un fragment de discours de François Mitterrand. Il s'agit, pour les participants, de séparer le document original de l'insert éventuel. Ils peuvent, à cette fin, s'appuyer sur un corpus de pseudo-documents annotés par les organisateurs.

Cette tâche se prête *a priori* à plusieurs approches :

– l'identification *non-supervisée* de segments thématiquement homogènes dans les pseudo-documents, problème pour lequel de multiples méthodes sont disponibles (voir, par exemple, (Hearst, 1997; Choi, 2000)), qui toutes essaient de tirer parti de l'organisation séquentielle du texte. Cette démarche est confortée par la méthodologie de constitution de la base de données, selon laquelle les insertions de discours de François Mitterrand traitent de thématiques différentes de celles des discours de Jacques Chirac. Cette stratégie a pour inconvénient de ne pas réellement exploiter les données de supervision disponibles.