Segmentation thématique par calcul de distance thématique

Alexandre Labadié*, Jacques Chauché*

* LIRMM, Université Montpellier 2
UMR 5506
161 rue Ada
34392 Montpellier Cedex 5 - France
alexandre.labadie@lirmm.fr,
jacques.chauche@lirmm.fr

Résumé. Dans cet article, nous présentons une approche de la segmentation thématique fondée sur une représentation en vecteurs sémantiques des phrases et des calculs de distance entre ces vecteurs. Les vecteurs sémantiques sont générés par le système SYGFRAN, un analyseur morpho-syntaxique et conceptuel de la langue française. La segmentation thématique s'effectue elle en recherchant des zones de transition au sein du texte grâce aux vecteurs sémantiques. L'évaluation de cette méthode s'est faite sur les données du défi DEFT'06.

1 Introduction

Le volume toujours plus important de textes rend l'exploitation de ces derniers par des méthodes automatiques de plus en plus complexes. Face à ce problème, la segmentation thématique offre la possibilité d'isoler dans un texte, des segments cohérents du point de vue de leur contenu informationnel. Ainsi, d'autres tâches telles que le résumé automatique ou la recherche d'information par exemple s'en trouve simplifiées. Mais l'on peut imaginer des tâches plus spécifiques telles que la création automatique de table des matières ou de plans à partir d'un gros volume de données non structurées. Nous présentons ici une approche originale de la segmentation thématique en nous appuyant sur les données du défi DEFT'06, Azé et al. (2006).

Pour son édition 2006, DEFT a fixé comme tâche de retrouver les différents segments thématiques d'un grand volume de textes. Trois catégories de textes nous ont été soumises :

- un ensemble de discours politiques.
- un ensemble d'articles de loi.
- un extrait d'un livre à teneur scientifique.

Chacune de ces catégories a été divisées en deux corpus distincts :

- Un corpus d'apprentissage, fourni au début du défi avec les segments thématiques étiquetés, afin d'entraîner nos méthodes.
- Un corpus de test, fourni à la fin du défi, sur lequel nous avons été évalués.

Un calcul de $Fscore$ sur les phrases frontières rapportées par les méthodes a permis l'évaluation des résultats. Les modalités du calcul du $Fscore$, et du couple rappel / précision qui lui est lié, dans le cadre de ce défi sont explicités par Azé et al. (2006).