1 Introduction

La méthode TextTiling[1] est un algorithme de détection automatique des changements de sujet au sein d'un texte. L'objectif de ce dernier est de permettre un découpage du texte motivé d'un point de vue sémantique.

C'est dans le cadre de mon stage de fin d'études, réalisé au sein du centre de recherche Multitel A.S.B.L. à Mons, que j'ai eu l'occasion de réaliser l'adaptation et l'implémentation en Perl de cet algorithme. Une des éventuelles utilisations possibles du programme ainsi réalisé était son intégration dans un système d'aide à la prise de décision, l'objectif étant alors de fournir à l'utilisateur la part d'information la plus pertinente possible par rapport à sa requête.

2 Principe général de la méthode TextTiling

Comme son nom l'indique, la méthode TextTiling propose un découpage du texte en unités continues qui ne se superposent pas [2]. L'idée générale dont découle cette méthode est qu'un texte traitant d'un sujet donné est en réalité divisé en une articulation de sous-sujets . Chacun de ces sous-sujets est évoqué grâce à un vocabulaire spécifique. Par conséquent, la transition entre ces sous-sujets correspond à un changement d'une partie relativement importante du vocabulaire.

Retrouver les frontières entre les diverses thématiques du texte équivaut dès lors à détecter les changements importants en ce qui concerne le vocabulaire employé. Les seuls éléments pris en compte par TextTiling pour détecter un changement de thématique sont les schémas de co-occurrence et de distribution lexicale. Les éventuels autres indices figurant dans le texte (sa disposition typographique, la référence des pronoms, etc.) ne sont pas pris en compte.

3 Les trois étapes de la méthode TextTiling

Les trois étapes essentielles qui constituent la méthode TextTiling sont :

[1] D'après Hearst 1997

[2] D'où l'expression *tiling*, qui signifie littéralement "carrelage".

1. La tokenisation du texte en unités de la taille d'une phrase. La tokenisation peut se définir comme l'opération de découpage du texte à analyser en groupes de mots, dont on a éliminé les éléments annexes, tels que par exemple les signes de ponctuation. Dans le cadre de TextTiling, cette tokenisation doit englober une opération de *lemmatisation*.
2. La deuxième étape majeure de la méthode TextTiling est constitué du calcul de deux types de scores, à savoir les scores de similarité lexicale et les scores de profondeur. Un score de similarité lexicale est calculé pour chaque intervalle entre deux portions contiguës de texte. Plus ce score sera élevé, plus le nombre d'unités lexicales communes entre les deux parties de texte sera importante.

   Quant au score de profondeur, il correspond à une brusque chute du score de similarité lexicale. En d'autres termes, si l'on représente sur un graphique l'ensemble des scores de similarité lexicale, ceux-ci vont dessiner une courbe plus ou moins accidentée. Les scores de profondeur correspondent aux vallées dessinées par cette courbe.

3. La détection des frontières entre les différents sous-sujets qui constituent le texte, grâce aux scores précédemment calculés.