Caractérisation et détection de thèmes

Pascale Sébillot - Projet TexMex - Irisa
sebillot@irisa.fr

Plusieurs travaux se sont intéressés à la détection automatique de thèmes en s'appuyant sur des indices linguistiques (*cf.* par exemple Litman et Passonneau) ou sur des notions telles que la cohésion lexicale (*cf.* par exemple Ferret et Grau, ou Hearst), certains d'entre eux réalisant simultanément la caractérisation de thèmes et la segmentation du discours. Pour notre part, nous ne nous intéressons pas à une structuration de la lecture séquentielle des textes, mais cherchons à déterminer une caractérisation permettant de connaître de manière immédiate le thème abordé dans un segment d'un corpus. Nous ne réalisons donc pas une analyse linéaire de l'intégralité de celui-ci. De plus, nous n'avons recours à aucune connaissance extérieure.

Fonctionnalités de l'outil

L'outil développé est un système de **caractérisation et détection de thèmes** dans un corpus textuel non spécialisé. Les principaux thèmes du corpus, dont la nature et le nombre sont initialement inconnus, sont **caractérisés par des ensembles de mots** symptomatiques de leur présence dans un texte, ou *mots-clés*, qui sont **extraits par un ensemble de traitements statistiques** exploitant la répartition des mots du corpus sur ses paragraphes. Les listes de mots extraites sont employées afin de **détecter la présence d'un thème** donné dans un paragraphe, par un simple critère de cooccurrence de mots-clés. Contrairement à de nombreux travaux entrepris dans ce domaine, le système ne réalise pas de découpage du texte en segments thématiques, mais une détection instantanée à l'échelle de paragraphes entiers. De plus, il ne fait usage d'aucune donnée auxiliaire, sémantique ou autre. Enfin, les classes de mots-clés obtenues constituent une caractérisation des thèmes du corpus aisément appréhendable par un lecteur humain, et rassemblent de nombreuses informations sur le corpus pris dans son ensemble.

Entrées-sorties et méthode

De manière un peu plus technique, l'outil prend **en entrée un corpus non spécialisé étiqueté morpho-syntaxiquement**. Un **algorithme de classification** ascendant hiérarchique est également utilisé : `Chavl` développé par I.-C. Lerman. **En sortie, les listes de mots** symptomatiques des principaux thèmes présents dans le corpus sont produites, listes qui peuvent ensuite être utilisées pour déterminer le thème abordé dans une unité textuelle.

L'obtention automatique des classes de mots se fait grâce à une suite de traitements statistiques :

1. Une première CAH classe les noms les plus fréquents du corpus en fonction de leur répartition sur ses paragraphes. L'algorithme de CAH manipule un tableau de contingence indiquant, pour chaque nom retenu, son nombre d'occurrences dans chacun des paragraphes.
2. Une seconde CAH classe les paragraphes du corpus en fonction de leur cohésion lexicale, c'est-à-dire des mots qu'ils partagent (avec un poids supplémentaire pour le partage de mots rares). Cette seconde classification sert tout d'abord à répondre à un besoin de densification de la matrice de répartition utilisée lors de la première classification. On passe en effet d'un tableau de contingence croisant les noms les plus fréquents et les numéros de paragraphes à un tableau croisant les mêmes noms et les classes de paragraphes obtenues, beaucoup moins creux. Elle est également utilisée pour définir une mesure de qualité des classes de noms et guider le choix des classes de mots-clés proposées par l'arbre de classification de noms, en apportant à celles-ci de légères modifications si nécessaire.
3. Plusieurs exécutions de ces deux classifications sont effectuées en prenant en entrée des extraits du corpus aléatoires distincts. L'objectif consiste à regrouper les mots les plus fréquemment rassemblés dans les diverses partitions obtenues. Pour ce faire, l'ensemble des partitions est synthétisé sous la forme d'un graphe valué dont les sommets correspondent aux mots classés et sont reliés entre eux par des arcs dont le poids est égal au nombre de fois où les mots ont été regroupés. Ce graphe permet d'extraire des noyaux restreints de mots, caractéristiques des thèmes du corpus.
4. Les noyaux sont ensuite étendus, en ajoutant à chacun d'eux les mots présentant, sur les paragraphes qu'il reconnaît comme abordant le thème qu'il caractérise, une fréquence anormalement élevée par rapport à leur fréquence moyenne sur le corpus.