Un outil de détection automatique de thèmes

Laurence Longo

Un outil de détection automatique de thèmes

Laurence Longo [1]

Laboratoire LiLPa (Linguistique Langues et Parole)
22 avenue René Descartes
67000 Strasbourg

Laurence.Longo@rbs.fr

MOTS-CLÉS : *recherche d'information, détection automatique de thèmes, cohésion, cohérence, marqueurs linguistiques.*

KEYWORDS: *information retrieval, theme detection, cohesion, coherence, discourse markers.*

Vu la quantité de documents numériques disponible sur le Web et la nécessité de mettre au point des techniques de recherche efficaces, les systèmes de recherche d'information font de plus en plus appel aux techniques de Traitement Automatique des Langues (TAL) qui exploitent les informations syntaxiques ou sémantiques, dans le but d'améliorer la qualité des résultats fournis par les moteurs de recherche, (Qristal, Intuition), (Illouz et *al*., 2000). Les moteurs de recherche actuels « plein texte » sélectionnent l'ensemble des documents contenant les mots-clés de la requête utilisateur. Nombreux sont les documents proposés à l'utilisateur qui ne comportent pas les informations attendues ; parfois même, des documents pertinents ne sont pas retrouvés par les moteurs. Ce manque de pertinence est dû à la méthode d'indexation par mots-clés, qui ne tient pas compte des propriétés linguistiques des textes (syntaxe, sens, genre etc.). Un aspect peu exploité à présent réside dans l'indexation automatique des documents par thèmes. Dans la lignée des méthodes hybrides existantes (Hernandez, 2004), nous allons combiner des méthodes statistiques à des méthodes linguistiques pour identifier automatiquement des thèmes et les proposer comme descripteurs de documents. Dans notre approche, les thèmes textuels constituent les sujets d'un texte, ou d'un fragment de document et sont posés comme agrégats des thèmes phrastiques (Goutsos, 1997). Ainsi, en plus des mots-clés, les documents seront décrits par leurs thèmes. Lors d'une recherche, les termes des requêtes utilisateur seront reliés aux thèmes déjà utilisés pour identifier les documents. Des thèmes associés pourront être proposés à l'utilisateur comme alternative et des documents associés au document consulté (*i.e.* comportant des thèmes proches) lui seront aussi proposés ; toujours dans une optique d'aide à la lecture et à la navigation.

1. Notre thèse est réalisée dans le cadre d'une convention CIFRE.

Ainsi, l'outil de détection de thèmes qui permettra d'améliorer l'indexation et la recherche par thèmes, suivra le schéma suivant. Le texte issu de documents de formats divers (PDF, Office ou HTML) va être extrait, puis le texte est segmenté en paragraphes par C99 (Choi et *al.*, 2001)[2], un algorithme statistique à base de cohésion lexicale. Afin de compléter ce découpage statistique et d'identifier les thèmes de chaque segment, nous appliquons des outils de TAL permettant de dégager des candidats-thèmes. Pour sélectionner les thèmes, nous nous appuyons sur des marqueurs linguistiques d'organisation textuelle de cohésion et de cohérence étudiés par des théories et modèles linguistiques (l'encadrement du discours (Charolles, 1997), les anaphores et les chaînes de référence (Schnedecker, 1997)). Nous combinons plusieurs catégories de marqueurs (marqueurs lexicaux de surface, marqueurs syntaxiques, chaînes de référence), séparément ou simultanément, pour identifier les thèmes de chaque segment. Nous proposons plusieurs critères d'attribution des thèmes décrivant le document et utilisés comme index : la fréquence des thèmes phrastiques, le rôle de ces thèmes dans la construction des chaînes de référence. La sortie de ce système est une liste de thèmes qui indexent chaque document. Notre système étant en cours d'implémentation, nous travaillons actuellement sur le module d'attribution de thèmes.

Bibliographie

Choi F. Y. Y., Wiemer-Hastings P., Moore J. « Latent semantic analysis for text segmentation », *Actes de NAACL'01*, 2001, p. 109-117.

Charolles M. « L'encadrement du discours : univers, champs, domaines et espaces », *Cahier de Recherche Linguistique* n° 6, 1997, p. 1-73.

Goutsos D. *Modeling Discourse Topic: sequential relations and strategies in expository text*, 1997, Norwood, N.J., Ablex Publishing Corporation.

Hernandez N. Description et détection automatique de structures de texte, Thèse de Doctorat, Université de Paris-Sud XI LIMSI/CNRS, 2004.

Illouz G., Habert B., Folch H., Fleury S., Heiden S., Lafon P., Prévost S. « TyPTex: Generic features for Text Profiler », in Content-Based Multimedia Information Access, *actes du colloque RIAO 2000*, Paris, 2000, vol. 2, p. 1526-1540.

Intuition http://www.sinequa.com/html-fr/fr-edition.oem.html

Qristal http://www.qristal.fr/

Schnedecker C. « Nom propre et chaînes de référence », *Recherches Linguistiques*, 21, 1997, Paris, Klincksieck.

2. Des méthodes de segmentation statistiques automatiques disponibles, notre choix s'est porté sur C99 car cet outil, de nombreuses fois évalué, s'avère être le plus performant.