Résumé de thèse

Rubrique préparée par Fiammetta Namer

Université Nancy2 de Nancy, UMR « ATILF »
Fiammetta.Namer@univ-nancy2.fr

Marion LAIGNELET (marion.laignelet@free.fr)

Titre : Analyse discursive pour le repérage automatique de segments obsolescents dans les documents encyclopédiques.

Mots-clés : analyse discursive, textes encyclopédiques, marqueurs linguistiques, segments obsolescents, repérage automatique, apprentissage automatique.

Title : *Discourse analysis for automatic location of obsolescent segments in encyclopaedic texts.*

Keywords : *discourse analysis, encyclopaedic texts, linguistic cues, obsolescent segments, automatic location, machine learning.*

en Sciences du langage, Université de Toulouse 2 – Le Mirail - CLLE-ERSS, départemment de Sciences du Langagesous la direction de Marie-Paule Péry-Woodley (Pr, CLLE-ERSS), et Ludovic Tanguy (MC, CLLE-ERSS). Soutenue le 25/09/2009.

Jury : Mme Marie-Paule Péry-Woodley (Pr, CLLE-ERSS, directrice), M. Ludovic Tanguy (MC, CLLE-ERSS, co-encadrant), Mme Agnès Tutin (MC, Université de Grenoble 3, présidente et examinatrice), Mme Lisbeth Degand (Pr, Université Catholique de Louvain, rapporteur), M. Patrice Enjalbert (Pr, Université de Caen, rapporteur), M. Claude de Loupy (ingénieur senior, entreprise Syllabs, Paris, examinateur).

Résumé : *La question de la mise à jour des documents se pose dans de nombreux domaines. Elle est centrale dans le domaine de l'édition encyclopédique : les ouvrages publiés doivent être continuellement vérifiés afin de ne pas mettre en avant des informations fausses ou altérées par le temps.*
Dans ce travail, nous proposons la mise en œuvre d'un prototype d'aide à la mise à jour : l'objectif visé est le repérage automatique de zones textuelles dans lesquelles l'information est potentiellement obsolescente.

Pour y répondre, nous proposons la prise en compte d'indices linguistiques sémantiquement variés et faisant appel à des niveaux d'analyses différents (fondés essentiellement sur la structure logique du texte et les titres). L'obsolescence étant un phénomène non linguistique pour lequel il n'existe pas d'outil rhétorique dédié, notre hypothèse est qu'il faut considérer les indices linguistiques et discursifs en termes de complémentarité et de combinaisons. La question est alors de savoir comment repérer automatiquement les combinaisons d'indices pertinentes sachant que le corpus en compte environ 150 différents et que la taille du corpus est conséquente (10 000 phrases).

Pour répondre à cette problématique, nous avons mis en place la méthodologie suivante. Sur un corpus annoté manuellement par des experts, nous projetons un repérage automatique d'indices linguistiques, qui ont pour particularité d'être discursifs et structurels. Un système d'apprentissage automatique par règles d'association est ensuite mis en place afin de faire émerger les configurations d'indices pertinentes dans les segments obsolescents caractérisés par les experts. Pour passer du modèle linguistique complexe au modèle statistique à deux dimensions, nous proposons un modèle de représentation des indices de discours intermédiaire qui nous permette notamment de gérer la variabilité du grain d'analyse mais également les relations entre des unités de discours différentes telles que les titres et les paragraphes.

Notre double finalité est remplie : nous proposons une description fine de l'obsolescence dans notre corpus de textes encyclopédiques (interprétation des statistiques, statistiques descriptives, analyse en composantes principales et apprentissage automatique) ainsi qu'un prototype logiciel d'aide à la mise à jour des textes.

Une double évaluation a été menée : par validation croisée sur le corpus d'apprentissage et par les experts sur un corpus de test. Les résultats sont encourageants pour une étude exploratoire sur le repérage automatique des segments d'obsolescence telle que celle que nous avons menée. Ils nous amènent également à reconsidérer la définition du segment d'obsolescence, sur la base des « découvertes » qui émergent du corpus, et dans l'interaction avec les besoins des experts concernant l'aide à la mise à jour. Ils montrent également les limites des annotations automatiques des indices linguistiques et discursifs. Le système peut et doit évoluer.

La reproductibilité de notre système doit finalement être évaluée ainsi que la pertinence et la réutilisabilité du modèle de représentation des données présenté.

Abstract**:** *The question of document updating arises in many areas. It is central to the field of encyclopedia publishing : encyclopedias must be constantly checked in order no to put forward wrong or time-altered information. In this study, we describe the implementation of a prototype of an aid to updating. Its aims is to automatically locate zones of text in which information might be obsolescent.*

The method we propose takes into account various linguistic and discursive cues calling on different levels of analysis. As obsolescence is a non-linguistic phenomenon for which no specific rhetorical tool exists, our hypothesis is that

linguistic and discursive cues must be considered in terms of complementarity and combinations.
Our corpus is first manually annotated by experts for zones of obsolescence.
We then apply automatic tagging of a large number of linguistic, discursive and structural cues onto the annotated corpus. A machine learning system is then implemented to bring out relevant cue configurations in the obsolescent segments characterized by the experts.
Both our objectives have been achieved : we propose a detailed description of obsolescence in our corpus of encyclopaedic texts as well as a prototype aid to updating.
A double evaluation was carried out : by cross validation on the corpus used for machine learning and by experts on a test corpus. Results are encouraging. They lead us to an evolution of the definition of obsolescent segments, first, on the basis of the "discoveries" emerging from our corpora and also through interaction with the needs of the experts with respect to an aid to updating. The results also show limits in the automatic tagging of the linguistic and discursive cues.
Finally, the reproducibility of our system must be evaluated as well as the relevance and reusability of the model of data representation.

URL où la thèse pourra être téléchargée :