

Sujet :

Analyse de texte pour l'aide à la numérisation de documents

Niveau : Bac + 5 - 3^{ème} année d'école d'ingénieur – Master

Durée du stage : 4 à 6 mois

Contexte

Le sujet du stage se situe dans le contexte de l'aide à la numérisation de documents en vue de leur indexation. Dans ce cadre, les outils de numérisation permettent, lors de la numérisation d'un document de transcrire l'image scannée du document complet en différentes zones d'image et de texte, ce dernier étant transcodé numériquement en suite de caractères grâce aux techniques de reconnaissance optique de caractères (OCR). La reconstruction d'un contenu textuel complet et cohérent à partir des zones de texte reste néanmoins un défi, en particulier parce que la mise en page du document est faite pour une interprétation visuelle par un œil humain et que l'ordre de lecture des zones de textes n'est pas toujours évidente.



Le document numérisé ci-contre contient deux textes:

- le premier est constitué d'un titre (1) et d'un contenu textuel composé dans l'ordre des zones:
 - 2
 - 5
 - 7
 - 9
 - 10
 - 3
 - 4
- le second est constitué d'un titre (11) et d'un contenu textuel composé dans l'ordre des zones:
 - 12
 - 13
- Les zones de textes 6 et 8 correspondent respectivement aux légendes des images 1 et 2 et ne sont donc pas rattachées aux deux textes.

Objectifs du stage

L'objectif du stage est de mettre au point les techniques et algorithmes pour déterminer l'ordre de lecture des blocs de texte reconnus par le système de numérisation. Ces algorithmes s'appuieront sur une analyse linguistique des zones de texte, pour déterminer différents critères d'ordonnement des blocs. On cherchera en particulier à exploiter:

- la cohérence syntaxique et sémantique: dans le cas de phrases coupées en fin de bloc, on cherchera à recoller une fin de phrase en début d'un autre bloc de façon à constituer une phrase qui soit grammaticalement cohérente et qui fasse sens ;
- la cohérence thématique: on s'appuiera sur des mesures de proximité thématique entre les



blocs (habituellement utilisées pour faire de la segmentation thématique de textes) pour déterminer les suites possibles d'un bloc et également repérer les ruptures thématiques fortes (pour déterminer les blocs n'appartenant pas aux mêmes textes) ;

- la cohérence discursive: de façon plus fine, on pourra également s'appuyer sur des techniques d'analyse du discours, permettant de structurer réthoriquement le texte par des relations de discours (développement, commentaires, explications etc.), qui permettront d'obtenir des mesures de cohésion discursive des textes.

Le laboratoire LIC2M du CEA LIST dispose d'une plate-forme modulaire de traitement des langues qui permet de réaliser l'analyse linguistique d'un texte, jusqu'au niveau syntaxique au moins. Le stagiaire s'appuiera sur cette plateforme pour l'analyse linguistique des zones de texte et devra déterminer, à partir de ces analyses, des mesures de cohésion entre les blocs de textes. On pourra éventuellement s'appuyer également sur des techniques d'apprentissage automatique pour certains des critères mis en œuvre.

Bibliographie :

Olivier Ferret, Approches endogène et exogène pour améliorer la segmentation thématique de documents , *Traitement Automatique des Langues* 2006 Volume 47 Numéro 2

Nicolas Hernandez, Brigitte Grau Analyse thématique du discours :segmentation, structuration, description et représentation, in *Proceedings of CIDE 2002*

Domaines de spécialité requis :

Traitement automatique des langues

Moyens informatiques mis en œuvre :

C++, Perl. Environnement Linux

Autres moyens mis en œuvre (expériences, méthodes d'analyses, autres...)

Centre : Fontenay-aux-Roses

Pôle ou Direction : DRT - LIST

Dépt/Service/Labo : DTSI/SRCI/LIC2M :

Laboratoire d'Ingénierie de la Connaissance Multimédia Multilingue

Web :

<http://www-list.cea.fr/>

Encadrement

Nom de l'ingénieur responsable : Romaric Besançon

Tél. : +33 1 46 54 80 16

E-mail : romaric.besancon@cea.fr

Fax : +33 1 46 54 75 80

Secrétariat : DTSI/SRCI

Tél. : +33 1 46 54 91 17

Nom du chef de laboratoire : Olivier Mesnard

Tél. : +33 1 46 54 73 38