

Analyse et structuration d'une requête pour l'interrogation de documents médicaux

Responsables :

Haïfa Zargayouna (email : haifa.zargayouna at lipn.univ-paris13.fr)

Sylvie Salotti (email : sylvie.salotti at lipn.univ-paris13.fr)

Contexte :

Le stage se déroulera au sein de l'équipe RCLN du LIPN et s'inscrit dans le cadre du projet de coopération avec le LIM&Bio (Laboratoire d'Informatique Médicale et Bioinformatique) de l'Université Paris 13. Ce projet vise à développer des méthodes originales d'accès aux documents médicaux.

Le stage porte sur le problème du passage d'une requête en langage naturel à une requête structurée exprimée dans un langage formel.

Ce problème correspond à une préoccupation actuelle dans le domaine de la Recherche d'Information Structurée (RIS) qui vise à exploiter à la fois la structure et le contenu des documents XML. En effet, la plupart des travaux de ce domaine proposent à l'utilisateur d'exprimer sa requête dans un langage formel ce qui suppose de maîtriser une syntaxe parfois complexe et d'avoir une connaissance préalable de la structure des documents interrogés.

La campagne d'évaluation INEX (INitiative for the Evaluation of XML Retrieval), qui a pour but d'évaluer la performance de différents systèmes de RIS, a initié en 2005 la tâche NLPX (Natural Language Processing for XML Information Retrieval) qui vise à intégrer les outils de TAL (Traitement Automatique de la Langue) pour répondre au mieux aux besoins des utilisateurs.

Le projet concernant l'accès aux documents médicaux porte en particulier sur des Guides de Bonne Pratique (GBP), documents qui décrivent des ensembles de recommandations médicales. GEM (Guideline Elements Model) est une DTD (Document Type Definition) définissant un modèle de structuration XML pour les GBP (<http://gem.med.yale.edu/default.htm>).

Objectif :

L'objectif du stage est de proposer un outil qui traduise une requête exprimée en langage naturel en une requête structurée exprimée dans un langage formel tel que Xpath qui permette d'exploiter la structure du GBP.

Les différentes étapes du travail à réaliser sont les suivantes :

- * Structurer un ou deux GBP en XML en utilisant la DTD GEM et les éléments de structuration fournis par l'équipe du LIM&Bio

- * Établir un état de l'art sur les outils de TAL qui existent pour définir quel niveau d'analyse serait nécessaire pour répondre à nos besoins.

- * Définir et implémenter un scénario d'analyse des requêtes en langage naturel et leur traduction en langage formel.

- * Expérimentation : tester la validité du scénario sur un ensemble de requêtes fournies par l'équipe du LIM&Bio.

- * Analyser les résultats et tester la validité des résultats sur un autre type de corpus (INEX par exemple).

Profil recherché :

Le candidat devra avoir de solides compétences informatiques, un intérêt pour le TAL. Une connaissance de XML, Xpath, etc. sera appréciée.

Conditions :

Stage de 6 mois, financé sur projet.