Proposition de stage GT DOING

Titre du stage : **Construction d'un graphe de connaissances à partir des relations extraites dans des cas cliniques**

Disciplines scientifiques : Traitement Automatique des Langues ; Apprentissage et Bases de Données

Partenaires académiques :

Encadrants : LIFO (Nicolas Hiot), LLL (Anne-Lyse Minard), LIFO (Mirian Halfeld Ferrari Alves), LIFAT (Agata Savary)

Type de stage : M2, stage de 6 mois, **financé par la fédération ICVL**

Projet dans lequel s'inscrit le stage :

Le groupe de travail DOING@DIAMS se concentre sur la question de la transformation des informations (obtenues dans des documents ou des bases de données plus au moins structurées) en connaissance. Il réunit des chercheurs des laboratoires LLL, LIFO et LIFAT, spécialisés en Traitement Automatique des Langues, Apprentissage et Bases de Données.

À partir d'un schéma représentant des connaissances expertes, nous cherchons à extraire les informations de textes permettant d'instancier le graphe. L'approche visée consiste en l'extraction d'entités du domaine et de relations entre entités, relations représentant les arcs du graphe. Les entités comme les relations devront être typées, normalisées et associées aux nœuds du graphe.

Les informations extraites seront ensuite transformées en base de connaissance. Elle devrait permettre d'assurer la qualité des informations, d'offrir des méthodes permettant l'interrogation et l'analyse des informations et d'offrir des mécanismes pour assurer l'évolution cohérente des connaissances.

Le stage proposé portera sur l'extraction des relations et l'instanciation de graphes, et s'inscrira dans la continuité d'un stage de M2 réalisé au 1er semestre 2020. Ce dernier a conduit au développement d'un système de reconnaissances des entités médicales ([1]) et de la réalisation d'une première étude de la problématique de l'extraction des relations.

Descriptif du stage :

Les objectifs du stage seront de continuer le développement d'une chaîne de traitements pour des textes du domaine médical en français (corpus de cas cliniques utilisé pour la campagne DEFT 2020) qui permettra d'extraire des relations entre les entités, les typer et de représenter ces informations sous forme de graphes grâce notamment à la normalisation des entités. L'extraction des relations reposera sur des règles (ou patrons) qui utiliseront l'analyse syntaxique (en constituants [2] ou en dépendances [3]) et/ou en rôles sémantiques [4]. Les perspectives à la suite du stage seront la généralisation du système à d'autres cas d'usage.

Nicolas Hiot, doctorant au LIFO, sera l'encadrant principal de ce stage. Le LLL et le LIFAT seront des personnes ressources pour la partie extraction d'information, en particulier Anne-Lyse Minard (LLL) sur la problématique de l'extraction de relations en domaine médical et Agata Savary du LIFAT sur les questions de grammaires locales. Mirian Halfeld Ferrari Alves du LIFO encadrera le stage pour la partie définition du schéma/graphe et leur instanciation en fonction des besoins pour la base de données.

Le stage sera décomposé en plusieurs étapes :

- Réalisation d'un état de l'art des systèmes d'extraction de relations dans le domaine général et dans le domaine médical.
- Annotation manuelle d'un sous-corpus pour l'évaluation du système et un affinage de la définition de la tâche.
- Développement d'une méthode pour l'extraction des relations entre les entités d'intérêts et leur typage.
- Développement d'une méthode pour représenter les relations extraites entre entités sous forme de graphes.

Compétences requises :

- Étudiants de master en TAL ou master en informatique avec un intérêt fort pour le TAL
- Bonne connaissance de python et des méthodes de TAL (parsing, text mining, etc.)
- Capacité de travail en équipe pluridisciplinaire

Références :

[1] Anne-Lyse Minard, Andréane Roques, Nicolas Hiot, Mirian Halfeld Ferrari Alves, Agata Savary. DOING@DEFT : cascade de CRF pour l'annotation d'entités cliniques imbriquées. DEFT 2020 (workshop de TALN 2020).

[2] Anne-Lyse Minard, Anne-Laure Ligozat, Brigitte Grau, Apport de la syntaxe pour l'extraction de relations en domaine médical. TALN 2011.

[3] Brahim Batouche, Claire Gardent and Anne Monceaux. Parsing Text into RDF graphs. SEPLN 2015.

[4] Piek Vossen, Rodrigo Agerri, Itziar Aldabe, Agata Cybulska, Marieke van Erp, Antske Fokkens, Egoitz Laparra, Anne-Lyse Minard, Alessio P. Aprosio, German Rigau, Marco Rospocher, and Roxane Segers. Newsreader: Using knowledge resources in a cross-lingual reading machine to generate more knowledge from massive streams of news. In *knowledge-based systems, elsevier*, 2016.

CANDIDATURES : envoyer un CV et les relevés de notes aux encadrants (via email)

Contacts : nhiot@ennov.com, anne-lyse.minard@univ-orleans.fr, mirian@univ-orleans.fr, agata.savary@univ-tours.fr