Projet Master 1 IC (2024-25)

Titre du sujet : **Éditeur web de graphes UNL**

(23/9/24)

Nom des encadrants : Christian Boitet
Email des encadrants : Christian.Boitet@imag.fr,
Téléphone(s) : +33 (04574) 21459, 0660051969

1. Présentation du sujet

Poursuivre et si possible finaliser un projet bien avancé (EdWebUNL) visant à réaliser un éditeur Web de « graphes UNL ».
Il s'agit en fait d'hypergraphes (un graphe peut contenir un ou plusieurs « scopes »). Chaque arc porte une relation sémantique codée sur 3 caractères (*agt* pour « agent », *gol* pour « goal », *plt* pour « place to », …). Chaque nœud porte un UW (universal word ou « lexème interlingue ») ainsi qu'une liste éventuellement vide de traits sémantico-pragmatiques (comme *def* pour « défini », *past* pour « passé », *entry* pour « point d'entrée dans un scope).
Depuis le début du projet international UNL (fin 1996), il y a eu plusieurs versions des spécifications des graphes UNL et des documents UNL, et il serait souhaitable que les outils informatiques autour d'UNL soient paramétrables par ces versions, ainsi que par les dictionnaires (ou bases lexicales) .
Un « texte UNL » est un fichier html dans lequel sont insérées des « balises UNL » entourant les graphes UNL, un par phrase du texte. Ainsi, <org>…</org> contient la phrase dans la langue origine (source), et <tgt lg= "fr">…</tgt> une phrase équivalente en français (par exemple, une sortie de traduction automatique (TA), ou une sortie de TA post-éditée).

Le projet précédent a mené à la construction d'un parseur de documents UNL, réalisé en javascript (PEG.js ou PEGGY.js), et un début d'éditeur. Il faut maintenant le compléter, et ajouter des fonctions permettant la vérification

- des relations sémantiques, en fonction de la liste donnée par la spécification
- des traits sémantiques ;
- des UW contenues dans le » volume lexical UNL actif ».

On spécifiera aussi des fonctionnalités intéressantes :

- Corriger un UW dans le document UNL ;
- Compléter le volume UNL actif avec les UW manquantes ;
- Inversement, remplacer un UW du document UNL par un UW du volume lexical actif courant.

Dans la première partie du projet (jusqu'à mars), il s'agira d'assimiler les techniques et outils utilisés dans le projet précédent, ainsi si possible que ce qui concerne UNL dans le projet RAPID UNseL. On utilisera pour cela le site web lingwariume.org et le site de Tetras-Libre (société grenobloise spécialisée en ontologies, pour faire bref).

Dans la seconde partie du projet (d'avril à mi-juin), on passera à la réalisation : développement (en intégrant la documentation aux programmes), tests unitaires sur des fichiers à fabriquer pour tester différents cas, tests sur les fichiers actuellement utilisés par le serveur d'Ariane-Y, et comparaison de performances.

2. Références

ANTLR : on se référera à https://www.antlr.org et aux compilateurs déjà développés pour Ariane-Y.
Langages : Java est supposé connu.
Générateurs de compilateurs : Ii faudra étudier et comparer quelques outils (ANTLR, PEGGY….
Outils : utilisation de git et de la forge du LIG.
Documentation avec Doxygen : on se référera à https://www.doxygen.nl.

3. Positionnement du sujet

- Indiquez le niveau d'innovation du sujet proposé

Très innovant	❒	❒	❒	☒	❒	Classique

- Indiquez la disponibilité de la documentation relative aux technologies à mettre en œuvre

Beaucoup	❒	☒	❒	❒	❒	Aucune

- Indiquez le niveau d'abstraction du sujet

Théorique	❒	❒	☒	❒	❒	Pratique

- Indiquez la quantité de développement à réaliser

Beaucoup	❒	❒	☒	❒	❒	Peu

- Indiquez le niveau de difficulté des algorithmes à mettre en œuvre

Difficile	❒	❒	☒	❒	❒	Facile

- Indiquez le niveau d'interaction avec d'autres composants logiciels

Ecosystème complexe	❒	❒	☒	❒	❒	Application seule

- Indiquez le nombre d'étudian(e)s souhaité(e)s pour le projet : 2

- Indiquez les langages et technologies à utiliser :
Java, Javascript, ANTLR, Doxygen

4. Encadrement

- Combien de temps pouvez-vous consacrer à l'encadrement de projets chaque semaine ?
Les 2 encadrants devraient être disponibles :
1ère partie : 2 ½ journées par semaine.
2ème partie : 4 à 5 ½ journées par semaine.

- Indiquez vos contraintes quant à l'encadrement
Pas de contraintes a priori.

- Indiquez vos contraintes quant au sujet proposé
Ce sujet implique un bon équilibre entre théorie et pratique. Il faut non seulement aimer développer, avec rigueur (suivre les préceptes et méthodes des cours de GI !), mais aussi avoir de la curiosité, en particulier pour s'initier aux arcanes de l'analyse syntaxique LL(*) mise en œuvre dans ANTLR par Terence Parr. ANTLR est un outil extrêmement puissant. Il sera utile de lire le livre de Terence Parr (voir sur le Web). Idem pour l'outil similaire écrit en Javascript.
Au terme de ce projet, les étudiants auront beaucoup progressé en GI, techniquement (Antlr) et méthodologiquement (Doxygen, Git).

A retourner à :
Damien.Pellier@univ-grenoble-alpes.fr