Annotation de textes structurés par des relations sémantiques

Lieu : **IRIT** - UT2J ou UPS, Equipe : MELODI
Contacts : Catherine.Comparot@irit.fr , Nathalie.Aussenac-Gilles@irit.fr (05 61 55 82 93)
Dates : février 2016 - juin 2016
STAGE de Master 2 R

Dans le cadre d'un projet ANR MOANO, une chaîne de traitements modulaire a été réalisée pour annoter automatiquement des documents tirés du web (des fiches) par des réseaux de concepts liés par des relations sémantiques. Les concepts à relier ayant été annotés, les propriétés qui les lient sont identifiées à partir du texte et d'une ontologie de domaine. Ainsi, à partir de l'indication "sol: frais et humide" dans une fiche décrivant la fougère, deux relations pour décrire le concept de plante ont par exemple été identifiées, <:Fougere :necessiteNatureSol :Frais> et <:Fougere :necessiteDrainage :Humide>. L'approche développée exploite la structure et les termes présents dans les textes pour identifier des concepts et des relations entre ces concepts. Les annotations de concepts sont actuellement représentées de façon ad-hoc en RDF (modèle de données de l'outil d'annotation TextViz développé dans le cadre d'un autre projet ANR), les relations sémantiques identifiées quant à elles ne sont pas mémorisées en tant qu'annotations, mais viennent enrichir l'ontologie au fur à mesure de leur découverte.

Le travail consistera à intégrer dans le processus actuel, l'adoption d'un modèle de données d'annotation adapté du standard OA, un format ouvert du W3C pour les concepts et les relations. Ainsi, les annotations pourront être publiées sous forme de LOD (Linked Open Data) et partagées par d'autres systèmes d'annotations. Il s'agira également d'évaluer l'approche globale. Ceci consistera à l'appliquer à un autre corpus pour mesurer notamment la qualité des annotations produites, la réutilisabilité des modules, sa portabilité ; des outils pour fournir des indicateurs seront développés.
Il s'agira enfin d'étudier l'apport du modèle de données d'annotation ouvert pour améliorer l'annotation sémantique. Différentes pistes seront envisagées : 1) la prise en compte dans l'annotation des relations de deux phénomènes linguistiques précis : la présence de négations et l'expression de préférences, 2) la validation manuelle des annotations sémantiques d'un document via l'affichage de son contenu augmenté par une représentation de type Knowledge Graph de ses annotations, 3) l'évaluation de l'apport des relations sémantiques pour interroger/rechercher des informations.

Compétences : IA, web sémantique et ontologies, base de traitement automatique des langues, programmation JAVA et PERL.

Durée du stage : 5 mois (entre mars et juillet 2016).
Rémunération : prime de stage (de l'ordre de **554,40€** /mois)

Références :

- Projet MOANO : http://moano.liuppa.univ-pau.fr/
- N. Aussenac-Gilles, M. Kamel, C. Comparot, D. Buscaldi (2013). Construction d'ontologies à partir de pages web structurées (regular paper). Dans : Journées Francophones d'Ingénierie des Connaissances (IC 2013), Lille, 03-05/07/2013, R. Troncy (Eds.), AFIA, p. 1-17, 2013.
- http://www.irit.fr/publis/MELODI/ic2013-Moano-Kamel-Aussenac-Buscaldi-Comparot.pdf
- M. Laignelet, M. Kamel, et N. Aussenac-Gilles (2011). Enrichir la notion de patron par la prise en compte de la structure textuelle - application à la construction d'ontologie. M. Lafourcade et V. Prince (Eds.), Traitement Automatique des Langues Naturelles (TALN 2011), Vol 2, pp. 267–272.
- Valarakos, A., G. Paliouras, Z. W. Ras, et G. Vouros (2004). Enhancing ontological knowledge through ontology population and enrichment. In Engineering Knowledge in the Age of the Semantic Web, 14th International Conference EKAW 2004, Volume 3257, pp. 144–156. Springer.