Ecole Doctorale STIC

Sciences & Technologies de l'Information et de la Communication

ED STIC - Proposition de Sujets de Thèse

pour la campagne d'Allocation de thèses 2015

Axe Sophi@Stic :	aucun
Titre du sujet :	Annotation sémantique et aide à l'analyse de textes anciens : extraction automatique de connaissances à partir de textes et raisonnement
Mention de thèse :	Informatique
HDR Directeur de thèse inscrit à l'ED STIC :	Andrea Tettamanzi

Co-encadrant de thèse éventuel :

Nom :	Faron Zucker
Prénom :	Catherine
Email :	faron@i3s.unice.fr
Téléphone :	04 92 96 51 50

Email de contact pour ce sujet :	andrea.tettamanzi@unice.fr
Laboratoire d'accueil :	I3S

Description du sujet :

Ce sujet s'inscrit dans un projet interdisciplinaire sur les humanités numériques mené en collaboration avec le CEPAM dans le cadre du GDRI Zoomathia, qui vise à explorer les possibilités que les méthodes de l'intelligence artificielle peuvent ouvrir pour l'étude de la transmission des savoirs à travers l'analyse de ressources matérielles, iconographiques et surtout textuelles. Cette exploration se concentre sur un cas d'étude spécifique, notamment l'histoire de la zoologie pré-moderne.

Une activité cruciale pour l'analyse des textes anciens est d'aller au-delà de cette numérisation et de les annoter par les chercheurs en sciences humaines. On peut distinguer deux types d'annotations :

Des annotations structurelles (de type syntaxique), classiquement en XML, selon le standard TEI,

dont le but est de mettre en évidence la division en parties et sous-parties du texte et les liens entre les textes ou fragments de textes. Ce travail d'annotation, qui doit aujourd'hui être effectué par des experts humains, est particulièrement fastidieux. La mise au point de techniques pour assister l'expert et automatiser, sinon complètement, au moins partiellement, cette activité serait particulièrement bienvenue.
Des annotations que l'on pourrait qualifier de « critiques » (de type sémantique), en ce qu'elles ressemblent, par leur nature et par leur but, à l'apparat critique des philologues. De fait, dans les projets tels que SourceEncyMe1 ou Ichtya2, l'objectif de ce travail d'annotation est de produire une édition scientifique du texte annoté (Buard, 2015). Lorsque plusieurs chercheurs sont impliqués dans le travail d'annotation sémantique du texte, ils s'entendent sur un « mode » ou « patron » d'annotation et produisent un résultat commun. Or, parmi les caractéristiques de ces annotations, il y a une certaine subjectivité ou dépendance du point de vue du chercheur qui les produit, le fait qu'elles peuvent être attachées à des fragments de texte de taille assez variable, qu'elles peuvent se chevaucher, qu'elles peuvent être utilisées comme notes de travail par un chercheur individuel ou comme un outil collaboratif de communication et coordination par une équipe de recherche. Dans la perspective d'une capitalisation et d'une réutilisation des annotations produites au-delà d'une édition scientifique, pour l'analyse critique des textes annotés, l'attribution d'une annotation à un chercheur est donc essentielle, ainsi que la coexistence dans la base de connaissances d'annotations issues de différents chercheurs avec des points de vue différents, et le chaînage en fils de discussion d'annotations qui sont des réponses à d'autres annotations. En outre, l'ensemble des annotations d'un texte donné est, par sa nature, ouvert, dans le sens que des nouvelles annotations peuvent toujours s'ajouter à celles existantes. Enfin, d'une certaine manière, ces annotations sont externes au texte, dans le sens que le texte ne dépend pas d'elles, alors qu'elles en dépendent et y sont liées par des ancres qui identifient des fragments du texte. Ce caractère libre des annotations critiques nécessite une manière d'identifier leur ancrage au texte indépendamment de sa structure (cf. Hellmann et al. 2012).

La thèse suivra trois directions de recherche complémentaires :
1. l'annotation semi-automatique et collaborative des textes dans la perspective de produire des annotations de type sémantique qui peuvent être réutilisées et sur lesquelles des raisonnements automatiques peuvent être mis en œuvre ;
2. l'extraction de connaissances ontologiques (concepts et relations) à partir de ces annotations sémantiques par un processus inductif (synthèse) ;
3. le raisonnement sur ces annotations et leur visualisation pour aider l'expert humain dans l'analyse des textes ((visual) data analytics).
L'objectif dans cette thèse est d'apporter des éléments de réponse aux questions suivantes :
- Comment permettre aux chercheurs en sciences humaines d'annoter de manière collaborative des textes anciens indépendamment de leur format et de sorte que ces annotations soient publiées et réutilisées ?
- Comment identifier des fragments de texte pour les référencer dans les données RDF ?
- Quelles techniques de traitement automatique du langage appliquer aux textes anciens pour en extraire directement des annotations sémantiques ?
- Quelles techniques de traitement automatique de texte appliquer aux annotations en texte libre

produites par les scientifiques, pour en extraire des annotations formalisées ?
- Comment induire automatiquement des concepts et des relations à partir de ces annotations ?
- Quels raisonnements mettre en œuvre pour assister l'expert humain dans l'analyse critique des textes ?

La thèse se déroulera au sein de l'équipe SPARKS, groupe Wimmics, dirigée par Andrea Tettamanzi spécialiste en apprentissage automatique et Catherine Faron Zucker spécialiste de la représentation des connaissances et du raisonnement sur le web sémantique, tous deux impliqués dans le GDRI Zoomathia.
Le plan de thèse envisagé suit les étapes suivantes :
1) Etat de l'art et proposition d'une méthode pour l'identification de fragments de textes
2) Proposition d'un modèle de données RDF permettant de représenter des annotations sémantiques d'un texte produites par différents utilisateurs et associées à des fragments qui peuvent « se chevaucher ».
3) Apprentissage automatique de connaissances
a) Etat de l'art sur le traitement automatique de textes courts et proposition d'une méthode d'apprentissage automatique de connaissances à partir des annotations en texte libre d'un texte étudié.
b) Etat de l'art et proposition d'une méthode d'extraction automatique d'annotations sémantiques à partir des textes étudiés.
c) Etat de l'art et proposition d'une méthode d'extraction automatique de connaissances ontologiques à partir des annotations sémantiques du texte étudié.
d) Couplage des méthodes d'apprentissage proposées
4) Etat de l'art et proposition d'une méthode de visualisation des données pertinentes (en parallèle de l'étape 3).

Références

S. Beldjoudi, H. Seridi, C. Faron-Zucker: Improving Tag-based Resource Recommendation with Association Rules on Folksonomies, SPIM 2011
A. Bouaziz, C. Dartigues-Pallez, C. da Costa Pereira, F. Precioso, P. Lloret: Short Text Classification Using Semantic Random Forest. DaWaK 2014
P.Y. Buard, Modélisation des sources anciennes et édition numérique, Thèse de doctorat de l'Université de Caen Basse Normandie, 2015
E. Cabrio, J. Cojan, A. Palmero Aprosio, F. Gandon, Natural language interaction with the web of data by mining its textual side, Intelligenza Artificiale, vol. 6, no. 2, pp. 121–133, 2012.
O. Corby and C. Faron Zucker, STTL - A SPARQL-based Transformation Language for RDF. WEBIST 2015
S. Hellmann, J. Lehmann, S. Auer. Linked-Data aware URI schemes for referencing text fragments. EKAW 2012
Ted Nelson. Literary Machines, Mindful Press, 1963.

English version: