Institut de Recherche
en Informatique
de Toulouse

Séminaire

Mercredi 6 Mars 2019
14h00 – 15h30
UT3 Paul Sabatier, IRIT, Auditorium J. Herbrand

Pascal CUXAC
INIST, Nancy

Désambiguïsation et alignement d'entités géographiques dans les textes scientifiques

Résumé : Dans le cadre de la phase 1 du projet ISTEX (https://www.istex.fr), des méthodes d'enrichissements ont été adaptées et appliquées concernant notamment la classification thématique, l'indexation, l'extraction d'entités nommées. Dans cette présentation je me focaliserai sur des entités nommées et plus spécialement sur des entités géographiques de type « placeName » désignant des noms de lieux géopolitiques ou administratifs (ville, région, pays, etc.). Ces entités ont été extraites dans une précédente étape grâce au programme Unitex-Cassys, mis en œuvre par le Laboratoire d'Informatique de tours (http://tln.li.univ-tours.fr/Tln_Istex.html).

Je présenterai une méthodologie de désambiguïsation et d'alignement automatique s'appuyant sur une approche par apprentissage automatique non supervisé : elle consiste en une représentation vectorielle, de type « plongement lexical » (words embeddings), avec une adaptation du modèle Skip-Gram utilisant une approche bayésienne non paramétrique pour “apprendre” plusieurs prototypes associés à un terme (Bartunov, Kondrashkin, Osokin, & Vetrov, 2016, http://proceedings.mlr.press/v51/bartunov16.pdf). Cela nous permet de désambiguïser les placeNames dans chaque documents ou ils se trouvent, et de les aligner avec la ressource geonames. Je commenterai les résultats obtenus dans le cadre du benchmark "SemEval-2019 : Task 12 - Toponym Resolution in Scientific Papers" https://competitions.codalab.org/competitions/19948

05 61 55 65 10
info@irit.fr
www.irit.fr

Organisé par
Département GD - Equipe IRIS