Stage M2 TAL : Prototypage d'une librairie Python pour l'extraction d'information

Laboratoire/Entreprise : Unité MaIAGE, INRAE, Université Paris-Saclay

Durée : 6 mois

Encadrants : Arnaud Ferré et Louise Deléger

Contact : arnaud.ferre@inrae.fr

Contexte

L'extraction d'information est le domaine du Traitement Automatique des Langues Naturelles visant à extraire et à structurer automatiquement des informations contenues dans de grandes quantités de textes. Une extraction commence classiquement par une tâche de reconnaissance d'entité, puis peut être suivie par une tâche de normalisation d'entité (parfois nommée "*entity linking/disambiguation"* ou "*concept normalization*") et/ou par une tâche d'extraction de relation.

L'équipe Bibliome de l'unité de recherche MaIAGE de INRAE/Université Paris-Saclay est spécialisée dans la recherche méthodologique en extraction d'information, notamment en domaines spécialisés. Elle développe également des solutions d'extraction pour des applications finalisées appliquées au domaine des sciences du vivant.

Sujet

Aujourd'hui, la grande majorité des méthodes d'extraction sont codées en langage Python. Bien que commencent à apparaître certaines librairies standards pour le traitement automatique des langues naturelles et qui contiennent leurs structures de données (ex : Stanza [1] ou spaCy [2]), celles-ci ne représentent souvent pas suffisamment les objets manipulés spécifiquement en extraction d'information. Par exemple, elles ne contiennent pas de classes explicites nommées "mention" ou "concept", basiques en normalisation d'entité, et bien qu'il existe une classe plus abstraite capable de représenter en particulier une mention, celle-ci ne peut pas être définie comme discontinue (ex : le groupe nominal "*liver and pancreatic cancer*" contient deux mentions distinctes dont la mention d'intérêt "*liver cancer*", laquelle ne peut être représentée de façon discontinue). En conséquence, la plupart des chercheurs qui développent de nouvelles méthodes s'appuient encore sur des structures ad hoc adaptées à leurs tâches, mais peu partageables et posant même des questions en termes de reproductibilité.

Nous faisons l'hypothèse qu'une librairie standard définissant une structure de données plus spécifique, c'est-à-dire plus proche des besoins des méthodologistes en extraction d'information, permettrait une meilleure reproductibilité, une facilité de prise en main, et un gain de temps de développement et d'intégration des méthodes.

La/le stagiaire devra développer un prototype de librairie Python définissant des classes d'objets adaptées aux besoins des méthodologistes pour les tâches de reconnaissance et normalisation d'entité. Un premier travail de comparaison avec au moins une des librairies standards devra être mené. Si cela est pertinent, la librairie pourra être développée comme une extension d'une de ces librairies standards. Des méthodes de reconnaissance et de normalisation et des jeux de données d'évaluation seront mis à disposition pour permettre de mettre en place un cadre de développement expérimental. Ce travail passera par le développement de *parseurs* qui iront parcourir, analyser et extraire les éléments des fichiers de jeux de données (de différents formats) pour les instancier dans un programme grâce aux structures de la librairie développée. Dans un second temps, ce travail pourra être dérivé à l'extraction de relation.

Le stagiaire aura accès à un ordinateur fixe, aux serveurs de calculs du laboratoire, et, au besoin, à des infrastructures de calcul haute performance (ex : Lab-IA).

Profil du candidat

Etre formé(e) ou expérimenté(e) en traitement automatique des langues naturelles ou plus particulièrement en extraction d'information.

Autonome en programmation Python, notamment orientée objet.

Formation et compétences requises

Master 2 / dernière année d'école d'ingénieur en informatique, linguistique ou TAL. Ouvert à d'autres spécialités (ex : bioinformatique) selon expérience.

Adresse d'emploi

Centre de recherche INRAE de Jouy-en-Josas (78)

Références

[1] Qi, Peng, et al. "Stanza: A Python Natural Language Processing Toolkit for Many Human Languages." Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations. 2020.

[2] Honnibal, Matthew, and Ines Montani. "spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing." To appear 7.1 (2017): 411-420.