Construction d’un corpus biomédical annoté et évaluation des systèmes de reconnaissance d’entités nommées dans un domaine de spécialité

De plus en plus de ressources d’articles scientifiques fournissent en accès libre des résumés ou/et des publications entières, plus spécifiquement dans le domaine biomédical. Ces articles représentent une importante source d’information qui peut être exploitée, à des fins d’extraction d’information. Nous nous intéressons aux articles traitant le microbiote intestinal, provenant essentiellement du système Pubtator Central (PTC). Il s’agit d’un système Web qui inclut approximativement 3 millions d'articles entiers et 30 millions de résumés issus du domaine biomédical. Le but est d’identifier les noms d’espèces, de gènes et de maladies mentionnés dans ces derniers.

La reconnaissance d’entités nommées et l’extraction de relations sont des sous-tâches de l’extraction d’information, dans le domaine du traitement automatique des langues. Elle consiste à traiter des documents structurés ou non structurés afin, d’une part, d’identifier des mots ou des groupe de mots puis de les classer dans des catégories bien définies et, d’autre part, de détecter l’existence d’une relation entre 2 ou plusieurs entités nommées puis de la classer dans une catégorie bien définie. Différentes approches à base d’apprentissage automatique peuvent être utilisées, pour ce faire : des approches supervisées, semi-supervisées ou encore non supervisées. Nous faisons principalement appel aux modèles d’apprentissage profond, les Transformers, s’appuyant sur les plongements lexicaux contextualisés, tel que le modèle BERT et ses différentes variantes entraînées sur des données biomédicales, ainsi que la librairie scispaCy. Il s’agit de modèles nécessitant une importante quantité de données annotées. Pour cette raison, une phase d’annotation sera requise pour compléter le corpus actuellement en cours de construction.

Les objectifs de ce Travail d’Etude et de Recherche (TER) se définissent comme suit :

- Evaluation de la pertinence des articles sur le microbiote intestinal, par rapport à la requête utilisée pour les sélectionner ;
- Comparaison des modèles existants de reconnaissance d’entités nommées pour l’annotation des entités nommées du domaine biomédical sur les articles sélectionnés ;
- Annotation de relations entre entités nommées à l’aide d’une ontologie pré-établie ;
- Amélioration du corpus et annotation des entités minoritaires/importantes.

Profils des candidat.e.s

Une/deux personnes sont attendues, avec pour chacune une appétence pour le TAL et/ou l’apprentissage automatique. Des connaissances en programmation Python sont également demandées.

Encadrement

Solen Quiniou et Oumaima El Khettari (solen.quiniou@univ-nantes.fr, oumaima.el-khettari@univ-nantes.fr)
Lieu du TER : LS2N (Site de la FST)

Bibliographie

- Lee, Jinhyuk, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* 36.4. Pages 1234-1240. 2020.

- Tong, Yiqi, Yidong Chen, and Xiaodong Shi. A multi-task approach for improving biomedical named entity recognition by incorporating multi-granularity information. In *Proceedings of ACL-IJCNLP.* 2021.

- Mark Neumann, Daniel King, Iz Beltagy, and Waleed Ammar. ScispaCy: Fast and Robust Models for Biomedical Natural Language Processing. In *Proceedings of the BioNLP Workshop and Shared Task.* 2019.

- Chih-Hsuan Wei, Alexis Allot, Robert Leaman, and Zhiyong Lu. PubTator central: automated concept annotation for biomedical full text articles. *Nucleic Acids Research* 47(W1). Pages W587–W593. 2019.