Proposition de TER pour la formation
Master Machine Learning pour la Science des Données - Université Paris Cité

Information concernant l'encadrant

Encadrant(s) : Séverine Affeldt (*MCF*), Université Paris Cité, Centre Borelli UMR 9010
Email : severine.affeldt@u-paris.fr

Description générale du projet

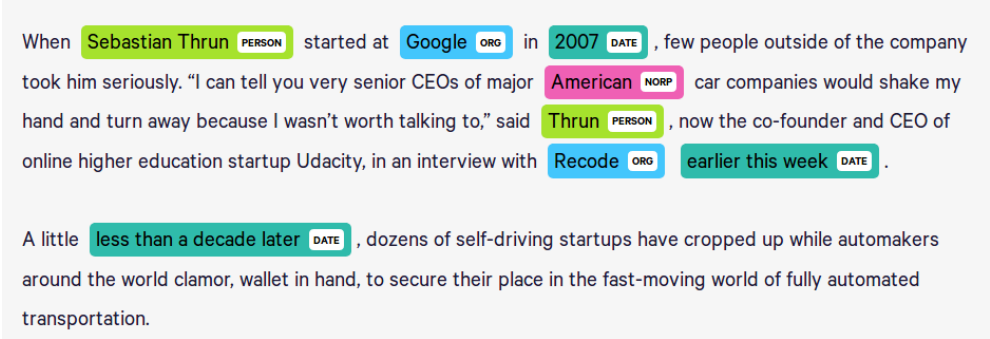
Intitulé du projet :

– Exploration de documents biomédicaux par reconnaissance d'entités nommées –

Contexte:

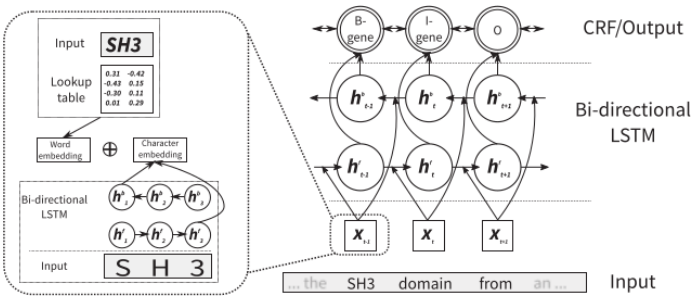
La détection d'entités nommées est un problème classique en NLP (Natural Language Processing). Une entité est une partie d'un texte (eg., mot, expression). Le fait d'extraire, à partir d'un large corpus de documents, des entités et de leurs attribuer des catégories (eg., localisation, organisation) constitue la tâche de détection d'entités nommées (NER; Named Entity Recognition) (Fig. 1).

Exemple de résultat pour une approche de NER – Fig.1



Dans un contexte biomedical, les approches NER sont importantes pour l'extraction d'informations concernant des entités biologiques (eg., maladie, gène, produit chimique) et leurs mise en relation à partir de grandes quantités de documents. Les premières approches de NER ayant atteint de très bons résultats comportaient généralement une étape basée sur un modèle *discriminatif* de champ aléatoire conditionnel (CRFs; Conditional Random Fields), et une étape exploitant un réseau profond de type LSTM (Long Short Term memory) (Fig. 2).

Exemple d'architecture Bi-LSTM + CRF – Fig.2



Les CRFs permettent de prendre en compte les interactions entre des variables proches et sont donc performants pour d'exploitation de séquences, comme c'est le cas pour le langage naturel. Un LSTM est un cas particulier de réseau de neurones récurrents (RNN; Recurrent Neural Network) qui permet le traitement de séquence de données.

Objectifs:

Ce projet de TER s'inscrit dans la continuité de travaux en cours dans l'équipe de recherche, notamment les approches de co-clustering RBDCo (*Regularized Bi-directional Co-clustering*) et EBCO (Ensemble Block Co-clustering), pour le co-partitionnement de mots et de documents. RBDCo permet d'incorporer facilement une information supplémentaire de similarité entre les mots et/ou entre les documents pour améliorer le co-clustering de documents et de termes. EBCO propose une approche ensemble permettant d'intégrer des multiples co-clustering en un co-clustering consensus.

Récemment, un prototype d'interface web intégrant l'approche RBDCo, nommée CORPEX, a été proposé. L'objectif de CORPEX est l'exploration et l'analyse de corpus biomédicaux. Afin de guider l'utilisateur dans l'exploration des documents et thématiques, nous proposons d'augmenter les résultats de co-clustering (eg., RBDCo, EBDCO) à l'aide d'outils de NER. Il s'agit notamment d'entraîner des modèles pour la reconnaissances d'entités relatives aux maladies, symptômes, facteurs environnementaux et gènes. Les résultats de NER peuvent également être exploités en amont de la tâche de co-clustering afin d'améliorer les résultats.

Travail à réaliser:

Parmi les implémentations disponibles en ligne, deux librairies proposent des approches intéressantes pour le NER: spaCy et FLAIR. Elles offrent notamment des extensions pour la gestion d'informations biomédicales.

Les fonctionnalités à réaliser:

1. Proposition et réalisation d'une étude comparative entre spaCy et FLAIR pour le biomédical.
2. Exploitation de spaCy et/ou FLAIR pour la reconnaissance des entités *maladies*, *symptômes*, *facteurs environnementaux* et *gènes*.
3. Amélioration des outils de visualisation existants dans CORPEX, notamment afin de permettre la visualisation des entités reconnues (eg., histogramme de fréquence, word clouds, matrice interactive documents/termes).
4. Identification des relations entre les entités découvertes et présentation sous forme de modèles graphiques.

Quelques tutoriels NER spaCy:
- Clinical Named Entity Recognition Using spaCy
- How to train custom NER using spaCy
- Training a medication NER with spaCy

Quelques tutorials NER FLAIR:
- Training a biomedical NER with FLAIR
- HunFlair: an easy-to-use biomedical NER
- BioNerFlair
- Med-Flair: medical NER for disease and medications using FLAIR