

Reconnaissance d'entités nommées pour le français avec des architectures *Transformers*

Encadrant : francois.rolle@u-paris.fr

Contexte

Des modèles de langage du français ont tout récemment été publiés (CamemBERT, FlauBERT). C'est positif car même des outils avancés comme Spacy sont notoirement encore peu adaptés à traiter certaines langues dont le français. Il reste à évaluer ce que ces modèles peuvent vraiment apporter notamment ne ce qui concerne la reconnaissance d'entités nommées.

Travail à réaliser

Réaliser des expériences de reconnaissance d'entités nommées sur le jeu de données *Europeana Newspapers French* en utilisant tous les modèles disponibles pour cela dans la bibliothèque *Simple Transformers*. Les adresses à consulter sont :

- <https://github.com/ThilinaRajapakse/simpletransformers>
- <http://api.bnf.fr/europeana-newspapers-french-ner-ground-truth>

Rendu

Un code clair et commenté dans un notebook.

Un rapport décrivant les méthodes/métriques d'évaluation (basées sur segeval dont une description détaillée sera donnée) et présentant/comparant/discutant les résultats obtenus.

<https://pypi.org/project/segeval/>

Critères d'évaluation

Qualité du code. Qualité de la rédaction du rapport, de la discussion. Qualité de l'expérience.