Modèles multilingues pour la reconnaissance d'entités nommées

Xavier Tannier
xavier.tannier@sorbonne-universite.fr

Mots-clés : Traitement automatique des langues (NLP), apprentissage statistique, reconnaissance d'entités nommées, multilinguisme.

La reconnaissance d'entités nommées (REN) est une tâche du traitement automatique des langues (TAL, ou NLP) qui a pour objectif de détecter les mentions d'entités d'intérêt dans les textes : noms de personnes, de lieux, d'organisation dans le domaine général, ou par exemple noms de médicaments, de maladies, de traitements dans les documents médicaux.

Les approches récentes de REN s'appuient sur des "plongements lexicaux" (*embeddings*) contextuels tels que BERT, qui permettent une représentation vectorielle très fine des mots du texte. Certains modèles d'embeddings sont pré-entrainés sur des données en plusieurs langues, ce qui leur confère une dimension multilingue intéressante.

Par ailleurs, les données d'entraînement pour les REN sont massivement disponibles en anglais, mais beaucoup moins dans d'autres langues.

Le but du projet est d'explorer les capacités des modèles multilingues, associés à des données d'entraînement multilingues également, pour améliorer la reconnaissance d'entités nommées dans des langues moins dotées que l'anglais. Des études existent déjà pour le domaine général, nous tenterons de les reproduire pour des textes du domaine médical.

Ce projet vous permettra de :

- découvrir le TAL et la reconnaissance d'entités nommées
- manipuler des modèles d'entités nommées
- concevoir des expériences s'appuyant sur de l'apprentissage statistique
- maîtriser les protocoles d'évaluation permettant de conclure sur l'intérêt de certains choix

Comme beaucoup de projets de TAL, la composante de manipulation de corpus textuels est importante : prétraitement des documents, manipulation de différents formats de données, connaissance de la nature et du contenu des données.

Contact : xtannier@sorbonne-universite.fr