*__OFFRE DE STAGE **TRAITEMENT AUTOMATIQUE DES LANGUES / LINGUISTIQUE DE CORPUS** (niveau L3 ou Master)__*

Durée du stage : 4 à 6 mois (début : avril 2023)
Date limite de candidature : 24 mars 2023
Gratification : montant légal en vigueur (environ 600€ / mois).
Laboratoire d'accueil : Équipe PASTIS du ***LIASD*** (EA 4383), Université Paris 8, 2 rue de la liberté, 93 526 Saint-Denis

Sujet de stage

Caractérisation objective des domaines/genres/registres/modalités pour le TAL : le cas de la reconnaissance des entités nommées en français

Les performances annoncées des systèmes d'Intelligence Artificielle appliqués au Traitement Automatique des Langues (TAL) peuvent être mise à mal par la diversité des situations linguistiques dans lesquelles ils sont utilisés[1]. Les performances d'un modèle d'apprentissage entraîné sur une certaine catégorie de ressources textuelles (par exemple, du contenu journalistique, encyclopédique de type Wikipédia, etc.) sont souvent moindres lorsqu'il est appliqué sur des ressources d'un type différent (parole transcrite, prose littéraire, poésie, contenu scientifique, etc.). La remédiation à cette variation est généralement abordée en TAL du point de vue d'une étape d'adaptation des modèles sur ressources additionnelles (*transfer learning)*, sans étude des caractéristiques linguistiques qui permettraient d'expliquer ces variations de performances.

Or, on peut faire l'hypothèse que la difficulté pour un outil de réaliser une certaine tâche à partir d'un apprentissage automatique sur un corpus spécifique tient à la sur-représentation dans celui-ci d'un phénomène et/ou d'une sous-tâche particulièrement difficiles, ou encore que les notions de domaine (médical, juridique, scientifique…) / de genre (littéraire, informel,…) / de registre (soutenu, familier) ou enfin de modalité (écrite, orale) ont une pertinence classificatoire sur une tâche A mais pas sur une tâche B.

Ce stage sera consacré à l'**identification et à l'analyse des descripteurs linguistiques pertinents pour la classification textuelle dans le cadre de la reconnaissance automatique d'entités nommées (REN) en français**.

La REN est une tâche qui s'inscrit dans l'**extraction d'information** et s'attache à identifier dans un texte les segments faisant référence à des classes déterminées telles que les personnes, les lieux, les dates, les organisations, etc. Cette tâche comprend à la fois un volet de segmentation (identification des frontières de l'entité) et de classification (typage de l'entité). En fonction du schéma d'annotation choisi, le typage peut-être plus ou moins fin (voir les types et sous-types : https://t.ly/IJGz), et l'entité peut être structurée ou non (voir les composants https://t.ly/IJGz).

L'identification de ces entités est utile notamment au développement de moteurs de recherche performants, de systèmes de recommandation de contenus, ou à la pseudonymisation de documents contenant des données à caractère personnel. Cette tâche concerne naturellement une grande variété de contenus, qu'ils appartiennent au domaine journalistique, encyclopédique, médical, juridique, etc.

Le corpus FENEC issu du travail de (Millour *et Al.*, 2022) a permis de mieux caractériser les variations inégales des performances de différents outils de reconnaissance des entités nommées selon différents genres textuels définis *a priori.* Un des objectifs de ce stage est donc d'**améliorer l'interprétabilité de ces variations** en s'appuyant sur les corpus annotés en entités nommées disponibles. L'identification de descripteurs linguistiques pertinents permettra de mieux appréhender les compétences des différents outils et de mieux prédire quel modèle est le plus approprié dans une nouvelle situation d'usage.

Problématique

Les catégories textuelles traditionnelles (domaine/registre/genre/modalité) sont-elles pertinentes/optimales dans le cadre du développement et de l'adaptation d'outils de reconnaissance des entités nommées ?

Étapes et réalisation

Étant donné le caractère exploratoire du sujet, plusieurs cycles d'analyses et découvertes, comprenant les étapes ci-dessous, seront nécessaires :

---

1 Voir, par exemple, les actes de la conférences Robustesse et TAL (ROBUSTAL) https://hal-cnrs.archives-ouvertes.fr/LISN-TLP/hal-03853541v1.

1. Identification de descripteurs linguistiques pertinents pour la caractérisation des genres textuels vis-à-vis de la tâche d'annotation en entités nommées pour le français (voir (Fu *et Al.,* 2020)) ;
2. identification des corpus disponibles et calcul des descripteurs document par document ;
3. analyse des erreurs de différents outils de REN (eg : SpaCy[2] , CasEN (Maurel *et Al., 2011*)), mise en regard des traits identifiés ;
4. classification non supervisée de l'ensemble des textes présents dans les sous-corpus basée sur la distribution des descripteurs linguistiques : les catégories *a priori* sont-elles retrouvées ? De nouvelles catégories (par regroupement ou division) apparaissent-elles ?

La ré-utilisabilité des ressources et les programmes produits feront l'objet d'une documentation tout au long du stage. En fonction du profil et des appétences du ou de la stagiaire, les étapes pourront être approfondies ou adaptées. Le ou la stagiaire pourra par ailleurs être amené(e)s à participer à des manifestations scientifiques liées à la problématique étudiée.

Compétences particulières et formation requise

Ce stage s'adresse aux étudiant.e.s de licence 3 ou Master 1 ou 2 en traitement automatique des langues, mais également en informatique pour des personnes intéressées par la langue naturelle. Compétences attendues :

- Programmation Python pour l'analyse textuelle ;
- Connaissances d'outils TAL appréciées (outils fondés sur l'apprentissage, classifieurs, outils statistiques de lexicométrie) ;
- Curiosité linguistique et volonté de tester de nouvelles méthodes.

Candidature

L'étudiant.e sera accueilli.e dans les locaux de l'Université Paris 8 au sein du laboratoire LIASD. Ce travail fait par ailleurs l'objet d'une collaboration avec l'université de Tours.

Contacts:

Alice Millour, LIASD, am@up8.edu
Jean-Yves Antoine, LIFAT, jean-yves.antoine@univ-tours.fr
Yoann DUPONT, LATTICE, yoann.dupont@sorbonne-nouvelle.fr

References

Douglas Biber, Representativeness in Corpus Design, Literary and Linguistic Computing, Volume 8, Issue 4, 1993, Pages 243–257, https://doi.org/10.1093/llc/8.4.243

Guillaume Cleuziou and Céline Poudat. 2009. On the Impact of Lexical and Linguistic Features in Genre- and Domain-Based Categorization. In Proceedings of the 8th International Conference on Computational Linguistics and Intelligent Text Processing (CICLing '07). Springer-Verlag, Berlin, Heidelberg, 599–610. https://doi.org/10.1007/978-3-540-70939-8_53

Jinlan Fu, Pengfei Liu, and Graham Neubig. 2020. Interpretable Multi-dataset Evaluation for Named Entity Recognition. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 6058–6069, Online. Association for Computational Linguistics.

Honnibal, M., Montani, I., Van Landeghem, S., and Boyd, A. (2020). spaCy : Industrial-strength Natural Language Processing in Python.

Denis Maurel, Nathalie Friburger, Jean-Yves Antoine, Iris Eshkol-Taravella, and Damien Nouvel. 2011. Cascades de transducteurs autour de la reconnaissance des entités nommées [CasEN: a transducer cascade to recognize French Named Entities]. *Traitement Automatique des Langues*, 52(1):69–96.

Alice Millour, Yoann Dupont, Alexane Jouglar, Karën Fort. FENEC : un corpus à échantillons équilibrés pour l'évaluation des entités nommées en français. Conférence sur le Traitement Automatique des Langues Naturelles (TALN), Jun 2022, Avignon, France. ⟨hal-03680569⟩

2 Voir : https://spacy.io/models/fr.