AdminSet and AdminBERT: un jeu de données et un modèle de langue pré-entraîné pour explorer le dédale non structuré des données administratives françaises

Thomas Sebbag[1, 2] Solen Quiniou[1] Nicolas Stucky[1] Emmanuel Morin[1]
Nantes Université, École Centrale Nantes, CNRS, LS2N, UMR 6004, F-44000 Nantes, France
Explore, Carquefou, France
Emails : thomas.sebbag@univ-nantes.fr, solen.quiniou@univ-nantes.fr, nicolas.stucky@etu.univ-nantes.fr, emmanuel.morin@univ-nantes.fr

RÉSUMÉ

Les modèles de langue pré-entraînés (PLM) sont largement utilisés en traitement automatique du langage naturel (TALN), mais peu adaptés aux textes administratifs, souvent non standardisés et spécialisés. En France, l'absence de réglementation uniforme et l'hétérogénéité des sources compliquent le traitement des documents administratifs. Pour pallier ce problème, nous proposons AdminBERT, le premier modèle de langue pré-entraîné en français dédié aux documents administratifs. Nous évaluons AdminBERT sur la tâche de reconnaissance des entités nommées (REN), en le comparant à des modèles génériques, un grand modèle de langue (LLM) et une variante du modèle BERT. Nos résultats montrent qu'un pré-entraînement sur des textes administratifs améliore significativement la reconnaissance des entités nommées. Nous mettons à disposition AdminBERT, AdminSet (un corpus de pré-entraînement) et AdminSet-NER, le premier jeu de données annoté pour la REN sur des textes administratifs français.

ABSTRACT

AdminSet and AdminBERT : a Dataset and a Pre-trained Language Model to Explore the Unstructured Maze of French Administrative Documents

In recent years, Pre-trained Language Models (PLMs) have been widely used to analyze various documents, playing a crucial role in Natural Language Processing (NLP). However, administrative texts have rarely been used in information extraction tasks, even though this resource is available as open data in many countries. Most of these texts contain many specific domain terms. Moreover, especially in France, they are unstructured because many administrations produce them without a standardized framework. Due to this fact, current language models do not process these documents correctly. In this paper, we propose AdminBERT, the first French pre-trained language model for the administrative domain. Since interesting information in such texts corresponds to named entities and the relations between them, we compare this PLM with general domain language models, fine-tuned on the Named Entity Recognition (NER) task applied to administrative texts, as well as with a Large Language Model (LLM) and to a language model with an architecture different from the BERT one. We show that taking advantage of a PLM for French administrative data increases the performance in the administrative and general domains on these texts. We release Admin-BERT as well as AdminSet,

the pre-training corpus of administrative texts in French, and the subset AdminSet-NER, the first NER dataset consisting exclusively of administrative texts in French.

MOTS-CLÉS : documents administratifs, modèle de langue, reconnaissance d'entité nommées, corpus, français.

KEYWORDS: administrative documents, Language Model, Named Entity Recognition, corpus, French.

ARTICLE : **Accepté à** The 31st International Conference on Computational Linguistics (COLING 2025)

https://aclanthology.org/2025.coling-main.27/ .