

NER sur décisions judiciaires françaises : CamemBERT Judiciaire ou méthode ensembliste ?

Sid Ali Mahmoudi*, Charles Condevaux*, Bruno Mathis**, Guillaume Zambrano*,
Stéphane Mussard*

*Univ. Nîmes CHROME, Avenue du Dr Georges Salan, 30000 Nîmes
sid.mahmoudi@unimes.fr,
<https://chrome.unimes.fr>

**Centre Européen de Droit et d'Economie, ESSEC Business School,
3 Av. Bernard Hirsch, 95000 Cergy
bmathis@cegetel.net

Résumé. Nous étudions dans cet article les apports respectifs de différentes représentations de mots, de la méthode ensembliste et d'un *transformer* spécialisé que nous appelons CamemBERT judiciaire, sur la tâche de recherche d'entités nommées dans les décisions de justice françaises. Nous comparons les performances des modèles BiLSTM-CRFs entre eux, individuellement ou constitués en ensembles, et avec le modèle de Ngompé et al. (2019) pris comme référence à battre. Les résultats obtenus montrent une amélioration.

1 Introduction

La première expérimentation d'apprentissage automatique de décisions de justice a été conduite par Aletras et al. (2016). Elle a montré dans quelle mesure la machine pouvait parvenir à la même décision que les juges de la Cour européenne des droits de l'homme avec un algorithme SVM. Suárez et al. (2020) ont montré que depuis lors, les approches neuronales ont considérablement amélioré l'état de l'art dans le traitement automatique du langage naturel en général et dans la reconnaissance d'entités nommées (NER) en particulier.

La Cour de cassation française a été la première à appliquer la NER à des décisions de justice françaises. Barriere et Fourret (2019) ont, pour elle, annoté des arrêts avec 4 étiquettes, personne physique, personne morale, adresse et date de naissance, et ont fait entraîner ces annotations à un réseau de neurones. La reconnaissance d'entités fonctionne sur tout type de chaîne de caractères. Filtz et al. (2020) l'ont appliquée à des dates, et Fernandes et al. (2020) à des montants, dans des décisions judiciaires rédigées respectivement en allemand et en portugais. Mathis (2021) a étendu cette approche à de nombreuses autres entités liées à la procédure judiciaire, soit un total de 27 étiquettes, comme la juridiction et les dates de jugement. Sur un large ensemble de données (1706 décisions), il a évalué les champs aléatoires conditionnels (CRF), de Lafferty et al. (2001), SpaCy, de Honnibal et Montani (2017), Flair, d'Akbik et al. (2019), et DeLFT (Deep Learning For Text), de Lopez (2018).

La sortie en 2018 par Google de son modèle BERT, de Devlin et al. (2018), basé sur un vaste corpus généraliste, a été suivie de variantes sectorielles. Le modèle dit Legal BERT, de