NuNER: Pré-entraînement d'un encodeur pour la reconnaissance d'entités nommées avec des données annotées automatiquement

Sergei Bogdanov[1] Alexandre Constantin[2] Timothée Bernard[3] Benoît Crabbé[3]
Étienne Bernard[2]
(1) Twin (mais travail effectué précédemment au sein de NuMind)
(1) NuMind
(2) Université Paris Cité, CNRS, Laboratoire de linguistique formelle
etienne@numind.ai

RÉSUMÉ

Les grands modèles de langues (ou LLM, pour « large language models ») peuvent s'avérer très efficaces pour l'annotation de données, ouvrant la voie à de nouvelles approches pour développer des systèmes de traitement automatique des langues par apprentissage automatique. Dans cet article, nous détaillons l'utilisation d'un LLM dans le développement de NuNER, un modèle d'encodage du texte, compact et spécialisé dans la tâche de reconnaissance des entités nommées (ou NER, pour « named entity recognition »). NuNER fait ainsi partie de la famille des modèles de fondation spécialisés. L'intérêt de NuNER est qu'il ne nécessite que très peu de données d'affinage pour obtenir un système de NER performant, quel que soit le domaine cible. Nous montrons qu'en régime d'apprentissage avec peu d'exemples (« few-shot learning »), NuNER surpasse les principaux modèles de fondation de taille comparable et a des performances similaires à celles de modèles de bien plus grande taille. Nos expériences montrent que la taille du jeu de pré-entraînement mais aussi la diversité des types d'entités qui y occurrent jouent un rôle essentiel dans ces résultats. NuNER et l'ensemble de ses données d'entraînement sont disponibles sous licence libre MIT.

ABSTRACT

NuNER : Entity Recognition Encoder Pre-training via LLM-Annotated Data

Large language models (LLMs) can be highly effective for data annotation, paving the way for new approaches to developing natural language processing systems through machine learning. In this article, we detail the use of an LLM in the development of NuNER, a compact text encoding model specialised for the task of Named Entity Recognition (NER). NuNER thus belongs to the family of specialised foundation models. Its key advantage is that NuNER requires very little fine-tuning data to become a high-performing NER system, regardless of the target domain. We show that, in a few-shot learning setting, NuNER outperforms leading foundation models of comparable size and achieves performance on par with much larger models. Our experiments demonstrate that both the size of the pre-training dataset and the diversity of entity types it contains play an essential role in these results. NuNER and its dataset are open-sourced under the MIT licence.

MOTS-CLÉS : reconnaissance d'entités nommées, annotation, apprentissage avec peu d'exemples, extraction de relation, apprentissage sans exemple, apprentissage de représentation, distillation, jeux de données pour le TAL.

KEYWORDS: named entity recognition, data labelling, few-shot learning, relation extraction, zero-shot learning, representation learning, distillation, NLP datasets.

ARTICLE : **Accepté à** EMNLP 2024.

L'article complet (Bogdanov *et al.*, 2024) est disponible au sein de l'anthologie ACL, le modèle et les données sont disponibles sur la plateforme Huggingface (`https://huggingface.co/numind`).

Références

BOGDANOV S., CONSTANTIN A., BERNARD T., CRABBÉ B. & BERNARD E. P. (2024). NuNER : Entity Recognition Encoder Pre-training via LLM-Annotated Data. In Y. AL-ONAIZAN, M. BANSAL & Y.-N. CHEN, Éds., *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, p. 11829–11841, Miami, Florida, USA : Association for Computational Linguistics. DOI : 10.18653/v1/2024.emnlp-main.660.