

Fine-tuning des Modèles de Langage Large (LLMs) pour l’alignement d’entités au sein des graphes de connaissances (GCs)

Bill Gates Happi Happi*, Géraud Fokou Pelap**
Danai Symeonidou*** Pierre Larmande*

*Université de Montpellier, 75 Av. Augustin Fliche, 34090 Montpellier, France,
{bill.happi+pierre.larmande}@ird.fr

**Université de Dschang, Ouest Cameroun, geraud.fokou@univ-dschang.org

***INRAE, SupAgro, UMR MISTEA, 2 place pierre viala 34060 Montpellier, France,
danai.symeonidou@inrae.fr

Résumé. La recherche d’entités similaires dans les graphes de connaissances a toujours été un défi complexe. L’arrivée des LLMs a ouvert de nouvelles perspectives, notamment grâce au fine-tuning qui permet à ces modèles de se spécialiser pour des tâches spécifiques. Cet article propose d’utiliser les modèles GPT-2 et BERT pour développer un modèle généralisé permettant de résoudre les problèmes d’alignement d’entités (AE) sur divers jeux de données. Un protocole basé sur le réseau de Kolmogorov-Arnold (KAN) est également présenté pour pallier les limites des LLMs en termes d’interprétabilité et de coût computationnel. Les résultats montrent que GPT-2 surpasse BERT et KAN, offrant de meilleures performances de score F1 pour les défis d’alignement d’entités. Cette approche offre une meilleure capture des similarités linguistiques, syntaxiques et sémantiques entre les entités.

1 Introduction

L’alignement d’entités (AE) est essentiel pour relier des entités équivalentes dans différents graphes de connaissances (GCs). Traditionnellement basé sur des règles (Zou et Özsu (2017)), l’AE s’est amélioré avec les techniques modernes d’embedding et de deep learning (Lu et al. (2023)). Les LLMs, grâce aux transformeurs (Vaswani et al. (2017)), offrent une interprétation automatique des descriptions d’entités, mais leur coût et leur interopérabilité posent des défis (Tan et al. (2024)). En parallèle, les réseaux KAN (Liu et al. (2024)), fournissent une alternative plus transparente pour capturer des relations complexes. Cet article explore la généralisation de l’AE via le fine-tuning des modèles de langage GPT-2 et BERT et l’entraînement from-scratch de réseaux KAN. Nous avons évalué ces approches sur 8 ensembles de données divers. GPT-2 et BERT ont d’abord été fine-tunés sur des jeux de données individuels, puis sur des données combinées, et finalement testés sur des ensembles de données inconnus pour évaluer leur capacité à généraliser. KAN a également été entraîné from-scratch pour comparer ses performances. Nos contributions concernent le Fine-tuning de GPT-2 et BERT pour l’AE ;