Actes de CORIA-TALN-RJCRI-RECITAL 2025
Marseille, France, 30 juin au 4 juillet 2025

Graphes, NER et LLMs pour la classification non supervisée de documents

Imed Keraghel[1, 2] Mohamed Nadif[1]
(1) Centre Borelli UMR 9010, Université Paris Cité, 75006 Paris, France
(2) Kernix Software, 75014 Paris, France

RÉSUMÉ

Les récents progrès en apprentissage automatique, notamment les modèles de langage de grande taille (LLMs) tels que BERT et GPT, offrent des plongements contextuels riches qui améliorent la représentation des textes. Cependant, les approches actuelles de *clustering* de documents négligent souvent les relations profondes entre entités nommées ainsi que le potentiel des représentations issues des LLMs. Cet article propose une nouvelle approche qui intègre la reconnaissance d'entités nommées (NER) et les *embeddings* de LLMs dans un cadre fondé sur les graphes pour le *clustering* de documents. La méthode construit un graphe dont les nœuds représentent les documents et dont les arêtes sont pondérées par la similarité entre entités nommées, le tout optimisé au moyen d'un réseau de neurones convolutifs sur graphes (GCN). Cela permet un regroupement plus efficace des documents sémantiquement proches. Les résultats expérimentaux indiquent que notre approche surpasse les méthodes traditionnelles basées sur la cooccurrence, en particulier pour les documents riches en entités nommées.

ABSTRACT

Graph-Convolutional Networks : Named Entity Recognition and Large Language Model Embedding in Document Clustering.

Recent advances in machine learning, particularly Large Language Models (LLMs) such as BERT and GPT, provide rich contextual embeddings that improve text representation. However, current document clustering approaches often overlook the deeper relationships between named entities and the potential of LLM embeddings. This paper proposes a novel approach that integrates Named Entity Recognition (NER) and LLM embeddings within a graph-based framework for document clustering. The method builds a graph with nodes representing documents and edges weighted by named entity similarity, optimized using a Graph Convolutional Network (GCN). This ensures a more effective grouping of semantically related documents. Experimental results show that our approach outperforms conventional co-occurrence-based methods in clustering, notably for documents rich in named entities.

MOTS-CLÉS : Clustering de documents, Entités nommées, LLMs, Graphes, Apprentissage de représentations.

KEYWORDS: Document clustering, Named entities, LLMs, Graphs, Representation learning.

ARTICLE : **Accepté à** ECIR 2025, disponible à l'adresse suivante : https://doi.org/10.1007/978-3-031-88711-6_6.

Actes des 32ème Conférence sur le Traitement Automatique des Langues Naturelles (TALN), volume 2 : traductions d'articles publiés, pages 30–30.