La contribution des LLM à l'extraction de relations dans le domaine financier

Mohamed Ettaleb*, Mouna Kamel*,**
Véronique Moriceau*, Nathalie Aussenac-Gilles*

*IRIT, Université de Toulouse, CNRS, Toulouse INP, UT3, Toulouse
**Espace-Dev, Université de Perpignan

Résumé. L'extraction de relations (RE) est une tâche clé en traitement du langage naturel, visant à identifier les relations sémantiques entre des entités dans un texte. Les méthodes traditionnelles supervisées entraînent des modèles pour annoter les entités et prédire leurs relations. Récemment, cette tâche a évolué vers un problème séquence-à-séquence, où les relations sont converties en chaînes cibles générées à partir du texte d'entrée. Les modèles de langage, de plus en plus utilisés dans ce domaine, ont permis des avancées notables avec divers niveaux de raffinement. L'objectif de l'étude présentée ici est d'évaluer l'apport des grands modèles de langue (LLM) dans la tâche d'extraction de relations dans un domaine spécifique (ici le domaine économique), par rapport à des modèles de langue plus petits. Pour ce faire, nous avons considéré comme baseline un modèle reposant sur l'architecture BERT et entraîné dans ce domaine, et quatre LLMs, à savoir FinGPT spécifique au domaine de la finance, et XLNet, ChatGLM2 et LLama3 qui sont généralistes. Tous ces modèles ont été évalués sur une même tâche d'extraction, avec, pour les LLM généralistes, des affinements par few-shot learning et fine-tuning. Les expériences ont montré que les meilleures performances en termes de F-score ont été obtenues avec des LLM affinés, Llama3 obtenant les meilleures performances.

1 Introduction

L'extraction de relations vise à identifier et classer les relations entre entités dans des textes. Dans des domaines de spécialité, cette tâche est plus compliquée en raison de la diversité et de la comlexité des expressions linguistiques, ainsi que par la spécificité de la terminologie. Des modèles aptes à gérer ambiguïtés et structures variées sont donc nécessaires. Au cours de la dernière décennie, l'apprentissage profond a transformé la tâche de RE. Des modèles pré-entraînés comme BERT (Devlin, 2018) ont montré des performances remarquables en RE sur des textes généralistes. Dans des domaines spécialisés, des modèles tels que GPT-FinRE (Rajpoot et Parikh, 2023) utilisent l'In-Context Learning (ICL) pour extraire des relations spécifiques mais leurs performances restent toutefois limitées face aux ambiguïtés et structures complexes. Concernant les grands modèles de langue (*Large Language Models* - LLM), bien que ces derniers marquent une avancée majeure pour l'extraction de relations dans des

contextes variés, des recherches ont montré que leur utilisation n'apporte pas de gains significatifs par rapport aux petits modèles, la tâche RE relevant d'un problème de classification (Lepagnol et al., 2024). Toutefois, des techniques d'affinement comme le few-shot learning et le fine-tuning permettent d'améliorer les performances des LLM dans des domaines spécifiques. Le few-shot learning est basé sur des invites simples, alors que le fine-tuning, plus coûteux, nécessite un jeu de données annoté et des ressources computationnelles importantes. Les principales questions que nous cherchons à aborder dans cet article sont les suivantes :

— Dans quelle mesure les LLM peuvent-ils surpasser les SLM (Small Language Model) pour l'extraction de relations dans un domaine spécifique, en l'occurrence le domaine économique ?
— L'affinement des LLM est-il efficace pour l'extraction de relations spécifiques à un domaine ?
— Les améliorations de performance obtenues grâce au fine-tuning des LLM justifient-elles les coûts associés ?

Pour répondre à ces questions, nous avons mené plusieurs expériences, chacune impliquant un modèle de langue appliqué au corpus CORE (Borchert et al., 2023), une ressource de haute qualité spécialement conçue pour l'extraction de relations économiques. Les modèles testés sont : un modèle BERT, FinGPT conçu pour l'économie, et trois LLM génériques (ChatGLM2, XLNet, LLama3), ajustés via les méthodes de few-shot learning et de fine-tuning. Ces modèles ont été sélectionnés car ce sont des modèles open-source qui peuvent être déployés localement, ce qui garantit un contrôle total sur les données et les processus d'entraînement tout en respectant les politiques de protection des données. En effet, dans le domaine économique, la confidentialité des données est essentielle, notamment pour les informations sensibles. L'utilisation d'APIs avec des LLM propriétaires pose des risques de sécurité car les données doivent être partagées avec des serveurs tiers.

2 État de l'art

Les méthodes d'extraction de relations (RE) ont évolué, passant de processus en plusieurs étapes impliquant la reconnaissance d'entités nommées et la classification des relations (Zeng et al., 2014) à des architectures basées sur les transformers réalisant une extraction de bout en bout (Wang et al., 2020). Les modèles de type sequence-to-sequence (seq2seq) ont encore amélioré les performances en RE (Cabot et Navigli, 2021). Certaines approches ciblent l'extraction d'une relation unique par paire d'entités dans des phrases courtes, tandis que d'autres traitent des textes longs pour identifier toutes les relations possibles entre plusieurs paires d'entités.
En TAL, l'extraction de relations au niveau des phrases se concentre sur des relations générales comme l'hyperonymie, à l'aide de jeux de données annotés tels que SemEval-2010 Task 8 (Hendrickx et al., 2019) et TACRED (Zhang et al., 2017). Les méthodes d'apprentissage profond ont permis des avancées, comme les modèles informés par des connaissances Khaldi et al. (2021), qui n'exigent ni entraînement supplémentaire ni alignement entité-vecteur pour intégrer des connaissances factuelles. Les modèles récemment ajustés pour le secteur économique, comme FinGPT (Wang et al., 2023) et Fin-LLaMA (Todt et al., 2023), se distinguent par leur optimisation pour l'économie. Cependant, les méthodes d'extraction de relations par phrase se limitent souvent à une seule relation par paire d'entités, même lorsque plusieurs relations ou des relations n-aires existent.

Ces dernières années, la recherche a beaucoup progressé en intégrant des ensembles de données financières avec des modèles basés sur GPT comme GPT-3 et GPT-4 pour améliorer les applications en TAL (Mann et al., 2020). Les méthodologies dominantes se divisent en deux catégories : la première repose sur l'ingénierie des prompts (White et al., 2023) avec des LLM open-source, utilisant leurs paramètres existants, et la seconde sur des méthodes de fine-tuning supervisé, comme Instruction Tuning (Ouyang et al., 2022).
L'instruction tuning est une approche récente qui améliore la généralisation des LLM en les fine-tunant sur une variété de tâches, souvent par le biais de démonstrations (Wang et al., 2022). Cela permet d'exploiter les connaissances acquises lors du pré-entraînement pour rendre les modèles plus adaptables à de nouvelles tâches. Plusieurs stratégies d'adaptation ont été développées pour rendre le fine-tuning plus flexible et efficace. Le prefix-tuning (Li et Liang, 2021) met à jour uniquement un petit segment au début du modèle, réduisant ainsi la charge computationnelle. Une autre approche, la Low-Rank Adaptation (LoRA) (Hu et al., 2021) utilise des matrices de faible rang qui peuvent être entraînées indépendamment, minimisant ainsi le risque de surapprentissage et les besoins en stockage.

3 Méthodologie d'évaluation des modèles de langue pour la tâche d'extraction de relations économiques

3.1 Description de la tâche

Étant donné une phrase $S = \{w_1, w_2, ..., w_n\}$, une entité E est une séquence contiguë de mots $E = \{w_i, w_{i+1}, ..., w_j\}$, où $i \leq j$. La tâche consiste à extraire des relations sous forme de triplets, chacun composé d'une entité E_1, une relation $r \in \mathcal{R}$, et une deuxième entité E_2, soit (E_1, r, E_2). Dans le cadre de l'extraction de relations dans le domaine de l'économie, cet article évalue plusieurs méthodes : (1) un modèle BERT entraîné pour identifier les relations entre entités ; (2) les techniques zéro-shot et few-shot appliquées aux LLM, évalués avec peu ou pas d'exemples ; et (3) le fine-tuning des LLM. L'objectif est de comparer ces approches en termes de performance et d'efficacité, afin de déterminer la stratégie optimale pour l'extraction de relations dans les textes économiques, tout en mettant en évidence leurs forces et limites.

3.2 Méthodologie

Nous considérons comme baseline un modèle BERT entrainé pour le domaine économique, un LLM affiné sur ce domaine FinGPT, et les LLM généralistes Llama3, ChatGLM, et XL-Net qui sont open-source (pour les raisons évoquées ci-dessus). Pour l'affinage, nous utilisons un ensemble de données spécifiques au domaine et issues du corpus CORE, un ensemble de prompts que nous avons défini pour les étapes d'affinage, ainsi qu'une méthode d'optimisation des LLM qui vise à réduire les coûts computationnels pendant le processus de fine-tuning.

3.2.1 Définition des prompts

Un prompt est une entrée textuelle utilisée pour guider ou déclencher la réponse du modèle. Le prompt fournit le contexte, les instructions ou la question que le modèle doit interpréter et traiter afin de générer une sortie appropriée (Lyu et al., 2024). Dans notre cas, le prompt

est composé de trois éléments clés : l'instruction, la phrase d'entrée et le format de sortie. L'instruction est définie comme suit : *"Quelle est la relation entre {E1} et {E2} dans le contexte du paragraphe d'entrée ? Choisissez une réponse parmi : {list_of_relations}"* pour guider le modèle vers la relation entre les entités. Le format de sortie est (*(E1, Relation, E2)*), assurant que les relations générées soient cohérentes et structurées. Les prompts sont alors générés à partir du dataset.

3.2.2 Méthodes d'affinage

Afin de réduire les coûts computationnels significatifs du fine-tuning des LLMs et de répondre aux limitations des tâches d'extraction de relations, une solution efficace est nécessaire. Nous employons la méthode d'adaptation PEFT (Mangrulkar et al., 2022), avec une utilisation spécifique de LoRA (Low-Rank Adaptation) (Hu et al., 2021) qui réduit considérablement le nombre de paramètres à ajuster tout en maintenant des performances élevées. PEFT est compatible avec une variété de LLMs open-source, dont Llama3, ChatGLM, et XLNet. Cette approche est censée améliorer la précision de l'extraction de relations et est largement applicable au domaine économique.

4 Ressources du domaine

Nous présentons ici les ressources utilisées et les modèles testés dans nos expériences.

4.1 Jeu de données

Nous avons utilisé le jeu de données CORE (Borchert et al., 2023), conçu pour l'extraction de relations du domaine de l'économie. Ce corpus est annoté manuellement par 12 types de relations. Contrairement aux annotaitons obtenues par supervision distante, celles de CORE couvrent diverses formes d'entités (noms propres, noms communs, pronoms) et se concentrent sur des entités comme les entreprises, marques et produits, rendant l'identification des relations plus complexe. Nous avons réparti ces données en 4000 instances pour l'ensemble d'entraînement et 708 pour l'ensemble de test.

4.2 Modèles de langue testés

Nous avons évalué cinq modèles open-source, facilement déployables localement, pour l'extraction de relations économiques au niveau des phrases :
XLNet (Extra-Long Transformer Network) : un modèle basé sur l'architecture Transformer, développé par Google, utilisant le Permutation Language Modeling (PLM) pour traiter différents ordres de mots. XLNet est entraîné sur plusieurs jeux de données, dont BooksCorpus, Wikipedia, Giga5, ClueWeb 2012-B, et Common Crawl.
ChatGLM : un modèle bilingue optimisé pour les tâches de questions-réponses et de dialogue en chinois et en anglais, basé sur le cadre General Language Model (GLM) avec 6,2 milliards de paramètres. Il a été entraîné sur 1,2 téraoctet de texte en anglais et 1,25 téraoctet de texte en chinois. Dans nos expériences, nous avons utilisé la version ChatGLM2-6B.
Llama3 : développé par Meta, Llama3 est une famille de LLMs comportant 8 ou 70 milliards

de paramètres, optimisée pour les tâches basées sur des instructions, et performante dans les cas de dialogue. Il est pré-entraîné sur plus de 15T de tokens collectés exclusivement à partir de sources publiques. Dans nos expériences, nous avons utilisé le modèle Llama3 avec 8 milliards de paramètres (Llama3-8B).
FinGPT : un modèle open-source conçu pour l'analyse et l'extraction d'informations dans le domaine financier. Il est entraîné sur des jeux de données comme l'analyse des sentiments sur des actualités et des tweets, ciblant les tâches spécifiques au domaine financier.
BizBERT : une version fine-tunée de BERT, spécialisée dans l'extraction de relations commerciales. Entraîné sur des jeux de données spécifiques tels que BizREL (Khaldi et al. (2021)), BizBERT se concentre sur les entités et relations commerciales.

5 Expérimentations et résultats

Cette section présente la configuration expérimentale, les résultats obtenus, ainsi que leur analyse. L'objectif est de répondre aux questions de recherche suivantes :

RQ1 : Les LLM surpassent-ils les SLM, et dans quelle mesure ? Pour cela, nous évaluons plusieurs modèles de tailles différentes et comparons leurs performances.

RQ2 : L'affinage des LLM est-il efficace pour l'extraction de relations spécifiques à un domaine ? Nous examinons ainsi si des techniques comme le N-shot learning ou le fine-tuning améliorent les performances de ces modèles.

RQ3 : Les améliorations de performance obtenues grâce au fine-tuning des LLM justifient-elles le coût engagé ? Il sagit de déterminer si les gains en précision pour extraire des relations compensent les ressources computationnelles élevées nécessaires au fine-tuning des LLM.

5.1 Configuration expérimentale

Pour répondre à ces questions de recherche, nous avons réalisé des expérimentations sur un jeu de données spécifique au domaine économique, le jeu de données CORE. Nous avons donc comparé les modèles BizBERT (modèle BERT pré-entrainé sur un corpus du domaine des entreprises et ré-entrainé sur le corpus CORE), XLNet, ChatGLM , Llama3, et FinGPT. Pour ajuster les LLM, ils ont été fine-tunés sur 8 époques avec un taux d'apprentissage de $1e-4$, une taille de lot de 4 et une accumulation de gradient de 8 étapes. Toutes les expériences ont été menées sur une NVIDIA RTX8000 (24 Go de RAM). Les performances des modèles ont été évaluées à l'aide des métriques de Précision, Rappel et F1-Score.

5.2 Évaluation des Performances

Nous avons cherché à comparer l'efficacité des LLM par rapport à des modèles plus petits et traditionnels, tels que les modèles basés sur BERT, pour évaluer leur capacité à s'adapter à des tâches spécifiques comme l'extraction de relations spécifiques à un domaine. Les résultats de notre évaluation, présentés dans le Tableau 1, montrent les performances des différents modèles sur le jeu de données test de CORE. Nous avons commencé par tester les modèles en utilisant des techniques de zero-shot et few-shot learning, où seulement trois exemples étaient inclus dans le prompt (dans le cas du few-shot). BizBERT a été réentraîné sur les données d'entraînement CORE, tandis que FinGPT a nécessité un fine-tuning du modèle BLOOM (Le Scao

et al., 2023) sur le même jeu de données. Les résultats des modèles fine-tunés sont également inclus pour comparaison. D'après le Tableau 1, les modèles de langue de grande taille

Méthode	Zero-shot	Few-shot	Fine-tuning	Réentraîné
BizBERT	indisponible	indisponible	indisponible	0.71
XLNet	0.54	0.58	0.76	indisponible
ChatGLM	0.56	0.59	0.78	indisponible
FinGPT	indisponible	indisponible	0.76	indisponible
Llama3	**0.69**	**0.70**	**0.80**	indisponible

TAB. 1 – *Comparaison des scores F1 des modèles sur le jeu de données CORE.*

comme Llama3 et ChatGLM surpassent systématiquement les autres modèles comme BizBERT, en particulier après un fine-tuning adapté à des tâches spécifiques comme l'extraction de relations dans le domaine économique. Le fine-tuning améliore significativement les performances, comme en témoigne l'augmentation des scores F1 de 0,69–0,70 en zero- et few-shot learning à 0,80 après le fine-tuning de Llama3.

5.3 Efficacité du fine-tuning des LLM

Pour valider l'efficacité du fine-tuning des LLM, nous avons mené des expériences en utilisant le jeu de données CORE. Nous avons fine-tuné Llama3 avec LoRA sur différentes proportions des données d'entraînement : 10%, 30%, 50% et 70%, et comparé les résultats avec ceux obtenus en utilisant l'ensemble complet des données d'entraînement. Comme illustré dans le

Configuration de Fine-Tuning	Précision	Rappel	F1 Score
Llama3 + 10% Données	0.75	0.72	0.73
Llama3 + 30% Données	0.78	0.74	0.75
Llama3 + 50% Données	0.80	0.75	0.77
Llama3 + 70% Données	0.81	0.77	0.78
Llama3 + Toutes les Données	**0.82**	**0.79**	**0.80**

TAB. 2 – *Impact des techniques de fine-tuning sur les performances des LLM en BRE.*

tableau 2, les performances du modèle s'améliorent significativement grâce au fine-tuning, même avec une petite portion des données. Par exemple, avec 30% des données d'entraînement, le score F1 atteint 0,75, surpassant déjà les performances obtenues avec moins de données. Les gains deviennent plus progressifs au-delà de 50%, avec un score F1 atteignant 0,77, et augmentant légèrement à 0,78 avec 70% des données.

6 Discussion et conclusion

Les résultats de nos expériences montrent que le fine-tuning des LLM est une stratégie très efficace pour améliorer les performances sur des tâches spécifiques à un domaine, comme l'extraction de relations dans le domaine de l'économie. Les modèles comme Llama3 surpassent

systématiquement les modèles plus petits basés sur BERT et montrent des gains significatifs de performance lorsqu'ils sont fine-tunés avec des données spécifiques au domaine. Une observation clé est que le fine-tuning, même sur une petite fraction des données disponibles (30-50%), génère des améliorations substantielles. Cependant, des augmentations supplémentaires de la quantité de données entraînent des rendements décroissants, suggérant qu'une performance optimale peut être atteinte sans utiliser l'ensemble complet des données. Cela met en évidence l'importance de la qualité des données plutôt que leur quantité brute. En conclusion, cette étude démontre que le fine-tuning des LLM pour des tâches spécifiques au domaine améliore non seulement les performances, mais constitue également une solution rentable et évolutive.

Références

Borchert, P., J. De Weerdt, K. Coussement, A. De Caigny, et M.-F. Moens (2023). CORE : A few-shot company relation classification dataset for robust domain adaptation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, doi: 10.18653/v1/2023.emnlp-main.722.

Cabot, P.-L. H. et R. Navigli (2021). Rebel : Relation extraction by end-to-end language generation. In *Findings of the Association for Computational Linguistics : EMNLP 2021*, pp. 2370–2381, doi: 10.18653/V1/2021.FINDINGS-EMNLP.204.

Devlin, J. (2018). Bert : Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv :1810.04805*.

Hendrickx, I., S. N. Kim, Z. Kozareva, P. Nakov, D. Ó. Séaghdha, S. Padó, M. Pennacchiotti, L. Romano, et S. Szpakowicz (2019). Semeval-2010 task 8 : Multi-way classification of semantic relations between pairs of nominals. *CoRR abs/1911.10422*.

Hu, E. J., Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, et W. Chen (2021). Lora : Low-rank adaptation of large language models. *CoRR abs/2106.09685*.

Khaldi, H., F. Benamara, A. Abdaoui, N. Aussenac-Gilles, et E. Kang (2021). Multilevel entity-informed business relation extraction. In *International Conference on Applications of Natural Language to Information Systems*, pp. 105–118. Springer.

Le Scao, T., A. Fan, C. Akiki, E. Pavlick, S. Ilić, D. Hesslow, R. Castagné, A. S. Luccioni, F. Yvon, M. Gallé, et al. (2023). BLOOM : A 176b-parameter open-access multilingual language model. *CoRR*, doi: 10.48550/ARXIV.2211.05100.

Lepagnol, P., T. Gerald, S. Ghannay, C. Servan, et S. Rosset (2024). Small language models are good too : An empirical study of zero-shot classification. *arXiv preprint arXiv :2404.11122*.

Li, X. L. et P. Liang (2021). Prefix-tuning : Optimizing continuous prompts for generation. *arXiv preprint arXiv :2101.00190*.

Lyu, K., H. Zhao, X. Gu, D. Yu, A. Goyal, et S. Arora (2024). Keeping llms aligned after fine-tuning : The crucial role of prompt templates. *CoRR abs/2402.18540*.

Mangrulkar, S., S. Gugger, L. Debut, Y. Belkada, S. Paul, et B. Bossan (2022). Peft : State-of-the-art parameter-efficient fine-tuning methods.

Mann, B., N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, et al. (2020). Language models are few-shot learners. *arXiv preprint*

arXiv :2005.14165 1.

Ouyang, L., J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, et al. (2022). Training language models to follow instructions with human feedback. *Advances in neural information processing systems 35*, 27730–27744.

Rajpoot, P. K. et A. Parikh (2023). Gpt-finre : In-context learning for financial relation extraction using large language models. *CoRR abs/2306.17519*, doi: 10.48550/ARXIV.2306.17519.

Todt, P. B. W., R. Babaei, et P. Babaei (2023). Fin-llama : Efficient finetuning of quantized llms for finance.

Wang, N., H. Yang, et C. D. Wang (2023). Fingpt : Instruction tuning benchmark for open-source large language models in financial datasets. *CoRR abs/2310.04793*, doi: 10.48550/ARXIV.2310.04793.

Wang, Y., S. Mishra, P. Alipoormolabashi, Y. Kordi, A. Mirzaei, A. Arunkumar, A. Ashok, A. S. Dhanasekaran, A. Naik, D. Stap, et al. (2022). Super-naturalinstructions : Generalization via declarative instructions on 1600+ nlp tasks. *arXiv preprint arXiv :2204.07705*.

Wang, Y., B. Yu, Y. Zhang, T. Liu, H. Zhu, et L. Sun (2020). Tplinker : Single-stage joint extraction of entities and relations through token pair linking. *arXiv preprint arXiv :2010.13415*.

White, J., Q. Fu, S. Hays, M. Sandborn, C. Olea, H. Gilbert, A. Elnashar, J. Spencer-Smith, et D. C. Schmidt (2023). A prompt pattern catalog to enhance prompt engineering with chatgpt. *CoRR abs/2302.11382*, doi: 10.48550/ARXIV.2302.11382.

Zeng, D., K. Liu, S. Lai, G. Zhou, et J. Zhao (2014). Relation classification via convolutional deep neural network. In *Proceedings of COLING 2014, the 25th international conference on computational linguistics : technical papers*, pp. 2335–2344.

Zhang, Y., V. Zhong, D. Chen, G. Angeli, et C. D. Manning (2017). Position-aware attention and supervised data improve slot filling. In *Conference on empirical methods in natural language processing*.

Summary

Relation extraction is a key task in NLP, aimed at identifying semantic relationships between entities in a text. This study evaluates the contribution of LLMs to relation extraction in the economic domain, comparing them to a domain-specific BERT model. Four LLMs were tested: FinGPT, XLNet, ChatGLM2, and Llama3, using techniques such as few-shot learning and fine-tuning. The results show that Llama3, fine-tuned for the task, achieves the best performance in terms of F-score, surpassing other LLMs and BERT, highlighting the potential of LLMs for specialized tasks.