Défi TextMine 2025 : Utilisation des Grands Modèles de Langue pour l'Extraction de Relations dans les Rapports de Renseignement

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

Défi TextMine 2025 : Utilisation des Grands Modèles de Langue pour l'Extraction de Relations dans les Rapports de Renseignement

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

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

Introduction

L'extraction de relations (RE) vise à identifier et caractériser les relations sémantiques entre des entités dans un texte, une tâche clé en traitement automatique du langage naturel (TALN). Les approches supervisées traditionnelles reposent sur l'annotation des entités suivie de la prédiction des relations entre elles. Récemment, les méthodes séquence-à-séquence ont simplifié ce processus en générant directement les relations sous forme de chaînes cibles. Les grands modèles de langage (LLMs) se distinguent par leur capacité à traiter efficacement ces tâches complexes. Dans le cadre du défi TextMine 2025, l'objectif est d'automatiser l'extraction de relations à partir de rapports complexes pour le renseignement et la défense. Ce défi offre une opportunité unique d'évaluer les performances des LLMs dans des scénarios réalistes. Nous proposons une approche utilisant le modèle Llama3 (Dubey et al., 2024) pour détecter et classer les relations entre paires d'entités dans un texte. Nous combinons la puissance des LLMs avec des étapes de filtrage préalable reposant sur les types d'entités et de relations. L'objectif est d'évaluer dans quelle mesure un LLM peut répondre aux besoins de l'extraction de relations dans des contextes complexes, tout en mettant en lumière ses limites et les défis à surmonter.

Approche proposée pour l'extraction des relations avec Llama3

Pour extraire toutes les relations possibles entre les paires d'entités dans un texte, notre méthode se décompose en plusieurs étapes décrites ci-dessous.

1. Génération de toutes les combinaisons de paires d'entités : Pour chaque texte, toutes les entités mentionnées sont identifiées et combinées en paires. Si l'ensemble d'entités est représenté par $E = \{e_1, e_2, e_3, \ldots, e_m\}$, les combinaisons générées incluent toutes les paires possibles, y compris les paires auto-référentielles, telles que (e_1, e_2), (e_1, e_3), jusqu'à (e_m, e_m). Cette étape garantit que toutes les interactions potentielles entre les entités sont couvertes.

2. Filtrage des paires selon les types d'entités et relations possibles : Les paires générées sont ensuite filtrées en utilisant des définitions des relations possibles données par le guide d'annotation, stockées dans un dictionnaire appelé `Relations_Definition`. Ce dictionnaire associe des relations spécifiques à des types d'entités, par exemple : une entité de type **Actor** peut être en relation *Is_Located_In* avec une entité de type **Place**. Chaque paire

est vérifiée pour s'assurer que les types des entités correspondent à une relation valide dans `Relations_Definition`. Si une paire respecte ces contraintes, elle est conservée ; sinon, elle est écartée. Ainsi, seules les paires d'entités valides restent dans la liste finale.
3. Identification des relations potentielles entre entités : Chaque paire d'entités retenue après filtrage est vérifiée dans le jeu de données annotées (gold standard). Si une relation annotée existe entre les deux entités, elle est associée à la paire. Sinon, on lui attribue l'étiquette `PAS_DE_RELATION`.
4. Génération du prompt : Un prompt personnalisé est généré pour chaque paire. Ce prompt inclut le texte contenant les entités, les types des entités, et une liste des relations possibles selon les types des entités (incluant toujours `PAS_DE_RELATION`). La sortie attendue pour ce prompt est la relation correcte, si elle est présente, ou `PAS_DE_RELATION`.
5. Entraînement du LLM : Les prompts générés sont utilisés pour fine-tuner le modèle Llama3. Chaque prompt comprend comme entrée le contexte du texte et les informations sur les entités. La sortie attendue est la relation correspondante ou `PAS_DE_RELATION`. Ce processus d'entraînement permet au modèle de comprendre les relations entre les entités en fonction de leur contexte et des définitions disponibles dans `Relations_Definition`.

Résultats et discussion

Nous avons divisé le jeu de données annotées, composé de 1200 textes, en 600 textes pour l'entraînement de Llama3, 200 pour la validation et 400 pour le test. Le corpus de validation respecte la même distribution des classes que le corpus d'entraînement afin d'assurer une évaluation équitable des performances. Le modèle obtient un score F1 macro de **0,61** sur le jeu de validation et de **0,38** sur le jeu de test, mettant en évidence des difficultés à généraliser certaines relations. Parmi les relations bien détectées sur le jeu de validation figurent *HAS_CONSEQUENCE* (0,94), *HAS_QUANTITY* (0,86), *IS_OF_NATIONALITY* (0,93), *IS_OF_SIZE* (0,95) et *GENDER_FEMALE* (0,93). En revanche des relations comme *WAS_CREATED_IN* (7 occurrences), *DIED_IN*(15 occurrences) et *IS_BORN_IN*(15 occurrences) ont obtenu des scores de F1 très faibles (entre 0 et 0,15), ce qui s'explique par leur faible fréquence dans le corpus d'entraînement. Un autre facteur ayant impacté les performances est la qualité des annotations. De nombreuses annotations contenaient des erreurs, comme nous l'avons constaté sur un sous-ensemble du corpus de validation. Par exemple, une relation *Is_Located_In* était annotée entre "volés" (verbe) et "Lisbonne" (nom), ce qui complique l'interprétation sémantique. Une mention plus explicite, telle que "articles volés" au lieu de "volés", aurait permis de mieux refléter la relation et d'améliorer la qualité de l'entraînement. Ces erreurs ont probablement perturbé l'apprentissage du modèle. Finalement, les LLM tels que Llama3 rencontrent des difficultés pour généraliser lorsque les données présentent un déséquilibre important entre les classes de relations.

Remerciements Ce travail a été financé par le projet ECLADATTA ANR-22-CE23-0020.

Références

Dubey, A., A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, et al. (2024). The llama 3 herd of models. *CoRR abs/2407.21783*, doi:10.48550/ARXIV.2407.21783.