

Normalisation à base de règles: une stratégie efficace pour l'extraction d'événements fondée sur des LLMs

Luca Dini*, Pierre Jourlin **

*SATT Sud-Est, Marseille
luca.dini@univ-avignon.fr

** Laboratoire Informatique d'Avignon, Avignon
pierre.jourlin@univ-avignon.fr

Résumé. Dans cet article, nous explorons l'intégration d'un traitement symbolique des sorties d'un LLM pour obtenir une extraction d'événements à haute granularité. Nous montrons que la faiblesse des LLM dans la production d'informations structurées, souvent soulignée dans la littérature, peut être surmontée en concevant une fonction d'appariement (hybridation) adaptée au domaine. Afin d'appuyer cette affirmation, nous comparons les résultats d'une méthode d'apprentissage en contexte avec notre approche hybride et nous montrons que cette dernière permet d'obtenir des résultats supérieurs (+6,3 %) sur un nouvel ensemble de données, de triplets sujet-prédicat-objet dans le domaine médical (681 triplets pour 200 phrases). Ce résultat est obtenu en laissant le LLM (Llama-3) libre de générer les types de prédicats avec lesquels il est le plus familier, et en appliquant a posteriori une fonction de normalisation. Outre l'amélioration de l'explicabilité et de la contrôlabilité de la sortie, l'intervention d'une telle fonction (qui a été mise en œuvre en cinq jours) permet de réduire de moitié les émissions de gaz à effet de serre induites par le traitement du corpus.

1 Introduction

L'un des principaux domaines dans lesquels les LLMs¹ sont employés est la transformation d'informations non structurées (« textes libres ») en informations structurées (bases de données ou graphes). Dans cet article, nous explorons une nouvelle méthode pour réaliser cette transformation, qui s'appuie sur l'intégration de LLMs et d'un traitement symbolique (hybridation).

L'objectif fonctionnel de cette recherche est d'extraire des informations de haute qualité, par exemple, sous la forme de graphes de connaissance, à partir de rapports de cas médicaux. Pour ce faire, nous construisons d'abord un corpus de 500 extraits de rapports médicaux² dans le domaine de la cardiologie.

1. *Large Language Models*, en français : grands modèles de langue

2. « Un rapport détaillé du diagnostic, du traitement et du suivi d'un patient individuel. Les rapports de cas patient contiennent également des informations démographiques sur le patient (par exemple, l'âge, le sexe, l'origine ethnique). » Définition du NIH.