Participation d'OppScience au challenge EvalLLM 2024 : une approche hybride applicable en contexte industriel

David Condaminet, Elias Limouni, Ferial Yahiaoui, Thibault Roy, Frédérik Bilhaut
OppScience, Pôle R&D, 14 avenue Trudaine, 75009, Paris, France
{dcondaminet,elimouni, fyahiaoui, troy, fbilhaut}@oppscience.com

RESUME

Cet article décrit la participation d'OppScience au challenge EvalLLM 2024, qui vise à évaluer les grands modèles de langues sur l'extraction d'entités nommées en français, à partir d'un corpus d'entraînement restreint favorisant les approches few-shots et les LLMs. Dans ce contexte, notre problématique industrielle est que les modèles génératifs demandent des ressources très importantes en phase d'inférence, alors que leur supériorité qualitative par rapport à l'ajustement supervisé de modèles pré-entraînés n'est pas avérée dans le cas général. Nous proposons une approche hybride combinant les avantages des modèles génératifs pour l'économie d'annotation, avec des modèles de type BERT économiquement viables pour le traitement de données massives, multilingues, et sensibles. Ce travail nous a permis la validité de principe de l'approche, qui se heurte toutefois à la subtilité du modèle d'annotation défini pour le challenge.

ABSTRACT

OppScience's Participation in the EvalLLM 2024 Challenge: A Hybrid Approach Applicable in an Industrial Context

This article describes the participation of OppScience in the EvalLLM 2024 challenge, which aims to evaluate large language models for named entity extraction in French, using a limited training corpus that favors few-shot approaches and LLMs. Our industrial challenge is that generative models require significant resources during the inference phase, while their qualitative superiority over supervised fine-tuning of pre-trained models is not generally proven. We propose a hybrid approach that combines the advantages of generative models for annotation efficiency with BERT models, more viable economically for processing large, multilingual, sensitive datasets. This work allowed us to validate the principle of our approach, though it faces challenges related to the complexity of the annotation model defined for the task.

MOTS-CLES : Extraction d'entités nommées, modèles neuronaux pré-entrainés, grands modèles de langage, applicabilité industrielle

KEYWORDS : Named Entity Recognition, Pre-trained Neural Models, Large Language Models, Industrial Viability

1 Introduction

Le challenge proposé dans le cadre de l'atelier EvalLLM 2024 vise l'évaluation de grands modèles de langues en français, plus spécifiquement sur une tâche d'extraction d'entités nommées et de marqueurs d'événements. La taille du corpus d'entraînement fourni aux participants, composé de quelques documents seulement, oriente *de facto* vers une approche de type *few-shots*, impliquant généralement la mise en œuvre de modèles pré-entraînés qui doivent ensuite être ajustés de façon supervisée (phase de *fine-tuning*) et/ou « invités » (via des *prompts*) à réaliser telle ou telle tâche spécifique. Différentes approches sont possibles dans ce contexte, qui vont de l'ajustement de modèles de type BERT à l'emploi de modèles génératifs ajustés et/ou guidés par le contexte. Nous renvoyons ici à (Xu et al., 2023) qui en dresse un panorama particulièrement complet.

Toutes ces méthodes ont pour point commun que les ressources demandées par l'ajustement et les inférences sont considérablement inférieures à celles qui sont nécessaires pour l'entraînement initial. Cependant, les modèles génératifs s'avèrent particulièrement exigeants en termes d'espace mémoire et de calculs, y compris en phase d'inférence, et ce même en recourant à des processeurs dédiés (GPU). Pourtant, leur supériorité qualitative sur les tâches d'extraction d'informations, dans leur définition classique, ne semble pas avérée à ce stade. Il en résulte que leur emploi au traitement de données massives se heurte, à l'heure actuelle, à un rapport peu favorable entre les coûts engendrés et la qualité d'extraction obtenue (quand elle est envisageable en termes de temps et de ressources matérielles). Ce qui motive l'émergence de travaux tels que (Zaratiana et al., 2023), dont l'objectif est précisément de réduire drastiquement l'effort de supervision sans subir le surcoût inhérent au paradigme génératif.

Ce constat est déterminant pour un éditeur logiciel tel qu'OppScience, dont la vocation est d'industrialiser des procédés d'IA multilingues et applicables à grande échelle avec des coûts raisonnables, a fortiori sur des données sensibles qui ne peuvent pas être déportées sur des infrastructures décentralisées. Cela nous a conduit à développer une approche hybride permettant de tirer parti à la fois des avantages des modèles génératifs en termes d'économie d'annotation durant la phase de mise au point, et de modèles plus frugaux pour la phase de production sur site.

C'est cette méthodologie, par ailleurs en cours d'intégration au sein de la plate-forme « Spectra », que nous avons choisi de mettre en œuvre dans le cadre de ce challenge. L'objectif n'est pas d'égaler les résultats qui seraient obtenus avec une approche *ad hoc*, mais de démontrer la viabilité économique d'une approche « sur étagère » applicable au traitement de données massives et sensibles. À ce sujet nous attirons l'attention sur le fait que la taille du corpus de test fourni dans le cadre de cet atelier n'est en rien comparable avec la volumétrie que nous adressons en situation réelle, ce qui suppose d'extrapoler la comparaison des coûts entre la phase de mise au point et de production.

2 Observations sur la tâche et le corpus

La tâche consiste en l'extraction d'entités d'intérêt pour des applications dans le domaine du renseignement (personnes, lieux, organisation, équipements, etc.) ainsi que des locutions appelées « amorces d'événements ». Les entités peuvent être imbriquées ou discontinues. L'extraction souhaitée doit mentionner pour chaque entité une classe et un ou plusieurs empans du texte d'origine spécifiés par les index de début et de fin, en préservant un certain nombre de mots grammaticaux qui composent le syntagme. Le guide d'annotation comprend 18 pages avec une

caractérisation très précise des classes selon des critères variés dont une partie significative est propre au domaine considéré. Les documents sont des bulletins d'information et des articles de blogs en français. Le corpus d'entraînement contient 5 documents pour 470 annotations réparties en 13 classes. Le corpus de test contient 24 documents.

Le catalogue d'entités correspond pour partie à la typologie que nous rencontrons usuellement sur des cas d'utilisation tels que la défense ou la sécurité intérieure. On notera toutefois que le guide d'annotation introduit de nombreuses nuances, plus subtiles que ce que l'on rencontre habituellement dans les tâches de NER, à tel point qu'il spécifie sous forme d'arbres de décision la méthode permettant de sélectionner la classe appropriée dans chaque cas, par exemple pour distinguer « organisation », « groupe », « unité militaire » et « fonction ». D'autres subtilités comme la distinction entre « lieu » et « site », ou encore la caractérisation des usages métonymiques des toponymes ajoutent au caractère particulièrement ambigu de la classification attendue.

Le concept d'amorce d'événement, s'il est défini avec une certaine précision dans le guide d'annotation, nous a semblé faire l'objet de quelques difficultés dans l'annotation du corpus. Nous avons en effet noté un certain nombre de cas où il nous semblait difficile d'expliquer le choix d'annoter certains éléments dont la portée aurait pu paraître anecdotique en termes de reconnaissance d'événements, ou au contraire d'en omettre d'autres qui auraient pu sembler pertinents.

De façon générale il nous a semblé que le modèle d'annotation reflète une tâche particulièrement ambitieuse dans un contexte *few-shot*, car la caractérisation des frontières entre les catégories repose sur des connaissances implicites du domaine qui peuvent être complexes à reconnaître de façon contextuelle, a fortiori sur la base de peu d'exemples. C'est évidemment un paramètre en faveur des capacités propres aux grands modèles de langue, à supposer que leur ajustement et/ou les instructions fournies puissent incorporer les critères nécessaires pour parvenir au niveau de discernement attendu.

3 Méthodologie

Notre méthodologie repose sur deux systèmes d'annotation complémentaires, que nous apellerons système de **production** et système d'**ajustement**.

Le système de production repose sur un modèle pré-entraîné et relativement économe en ressources, qui peut donc être mis en œuvre sur site et sur de grands volumes de données multilingues et potentiellement sensibles. Ceci est au prix d'un ajustement à partir d'un nombre suffisant d'exemples : de l'ordre de la centaine par classe. Ce nombre raisonnable pourrait permettre de qualifier cette partie du système dans la catégorie « few-shot », cependant cela demande un effort d'annotation sensiblement supérieur à ce qui est proposé dans le cadre du présent challenge. Il en va de même dans le cadre de nos cas d'usage industriels, où il est fréquent que les utilisateurs soient dans l'incapacité d'annoter un corpus de taille conséquente, a fortiori sur plusieurs langues.

Pour cette raison, nous nous appuyons en amont sur le système d'ajustement. Celui-ci repose sur des modèles génératifs, donc plus « coûteux », qui seront utilisés uniquement durant la phase de mise au point pour annoter un corpus destiné à l'ajustement du modèle de production. Cette phase peut être réalisée sur des données non sensibles, ne demande que très peu d'intervention humaine,

et ses coûts sont concentrés sur un nombre de documents très petit en comparaison de ce qui sera effectivement analysé en production.

En situation réelle, on aura compris que le système d'ajustement n'est pas utilisé en production. Néanmoins, dans le cadre de ce challenge, nous l'avons exploité dans sa fonction d'annotation afin de constituer un « run » en tant que tel.

3.1 Système de production

Il s'agit d'un système d'étiquetage de séquences basé sur une architecture de type transformer (à partir de BERT dans sa version « multilingual uncased »), implémentée sur les bibliothèques Flair (Akbik et al., 2019) et PyTorch. Intégré en standard dans notre produit, ce modèle a été entraîné sur des données en plusieurs langues, dont le français, dans le domaine de l'investigation et de la justice, anonymisées et annotées en internes, sur 10 classes dont 4 présentes dans le challenge : ORG, PER, LOC, DATE (néanmoins nous avons observé que la définition de ces classes et les formes textuelles couvertes diffèrent de celles spécifiées dans le guide d'annotation).

Pour ce challenge, l'objectif était d'ajuster notre modèle sur les 13 classes à partir des données d'apprentissage fournies qui ont été augmentées automatiquement avec la méthode décrite en 3.2.

Dans le but de maximiser la performance du modèle d'un point de vue qualitatif, une recherche en grille (*grid-search*) a été effectuée afin de déterminer les meilleurs hyperparamètres pour l'ajustement de notre modèle final. Cette recherche exhaustive a permis de tester au total 54 combinaisons différentes, en jouant sur la taille du batch, le taux d'apprentissage et le nombre d'époques, en apprenant sur les données générées puis en évaluant sur le gold. Le modèle ne devait ni s'adapter trop étroitement, ni trop peu, aux données d'entraînement, auquel cas il ne généraliserait pas suffisamment sur les données de test. Le meilleur modèle (batch de 8, taux d'apprentissage à 0.001, 250 époques) ajusté sur les données générées par la méthode décrite en 3.2 a obtenu les scores de 19,5% en macro-F1 et 24,1% en micro-F1 sur le corpus de référence.

Les limites de cette approche traditionnelle de la reconnaissance d'entités nommées par classification de tokens sont multiples. Ce système attend en entrée des séquences de jetons. Deux étapes de pré-traitement sont alors requises pour préparer correctement les données fournies sous la forme brute au modèle : la segmentation en phrases puis la tokenisation de celles-ci en mots. La qualité de ces transformations est alors cruciale, sur les données d'apprentissage pour le bon entraînement du modèle, mais aussi sur les nouvelles données pour l'inférence. Aussi, son format d'entrée de type CoNLL ne permet pas de représenter des entités qui se superposent ou qui sont imbriquées, cela entraînant alors une diminution de la couverture d'entités avant même l'apprentissage. De surcroît l'heuristique de choix de l'entité à conserver dans ce type de cas est discutable voire controversée.

Une autre perte d'information concerne les entités discontinues qui ne sont pas non plus représentables dans le format attendu par le système dont le schéma d'étiquetage est IOBES. Enfin, cette approche nécessite une augmentation du corpus d'apprentissage, produisant une sur-représentation de nouvelles données automatiquement annotées par LLM, ce qui crée une dépendance forte à la qualité du système LLM. Le risque d'ajuster notre modèle sur des données partiellement correctes, contenant du bruit et/ou du silence, est bien existant. Cela signifie également que la sortie de nos deux systèmes sera forcément corrélée.

Une autre approche non LLM, fondée sur une tâche de catégorisation de *spans* (sous-séquences de mots), communément utilisée pour la résolution de coréférences ou bien l'extraction de relations, pourrait être en effet plus appropriée, du moins dans des cas de figure présentant de telles entités.

3.2 Système d'ajustement

La fonction du système d'ajustement est de créer (ou augmenter) un corpus annoté qui sera utilisé pour ajuster le système de production. Ce système étant appliqué à une quantité de données réduites et non sensibles, il n'y a pas d'obstacle pratique à la mise en œuvre de modèles génératifs y compris sur des infrastructures déportées. L'approche retenue faisant l'objet d'une autre communication soumise dans le cadre de l'atelier (Yahiaoui et Limouni, 2024), nous nous contenterons ici d'en donner les principes généraux.

La technique repose d'abord sur des modules de génération d'entités spécifiques, telles que des numéros de téléphone, des entités financières, des numéros d'identité ou encore des entités liées à des véhicules. Chaque module crée des données synthétiques variées et réalistes, par exemple, des numéros de téléphone avec divers formats, des IBAN fictifs, ou encore des modèles de véhicules avec plaques d'immatriculation imaginaires. La coordination de ces modules garantit la couverture des types d'entités souhaitées, ainsi que leur diversité et leur précision.

Elle implique d'autre part un module génération selon des scénarios structurés, permettant à l'utilisateur de spécifier les caractéristiques des textes à produire à l'aide de formats déclaratifs. Ces scénarios pilotent le processus de génération des entités et la création de prompts destinés fournis à un LLM pour la génération de textes. L'approche inclut une étape de validation permettant de vérifier la pertinence des textes produits.

Pour estimer la qualité des textes générés, deux types de métriques sont utilisées : le taux de rejet et le taux de répétition sur la base de n-grammes. Le taux de rejet vérifie la présence effective des entités demandées dans les textes, tandis que le taux de répétition évalue la diversité des textes en détectant les séquences répétitives. Ces métriques ont permis d'optimiser la génération et de détecter les régressions lors des modifications des prompts. L'application de ces métriques a montré que la méthode permettait de générer des textes répondant efficacement aux besoins des utilisateurs, tout en assurant la diversité et la pertinence des données produites. L'évaluation qualitative et quantitative montre que les modules de génération d'entités offrent une diversité et une précision élevées, avec un f1-score pondéré moyen de 0,9 sur des données annotées automatiquement.

Dans le cadre de ce challenge, l'approche a été utilisée d'une part pour augmenter le corpus d'entraînement, et d'autre part en tant que module d'annotation pour fournir un run indépendant, afin de pouvoir comparer la méthode de production avec la partie LLM prise isolément, aussi bien en termes de qualité que de coûts.

4 Résultats

4.1 Analyse des coûts

La méthodologie que nous proposons ayant pour objectif de limiter les coûts en production, nous souhaitons en donner une analyse détaillée au même temps que nous répondons à la demande faite aux participants de tracer leur empreinte carbone.

Les coûts comptabilisés pour chacune des phases sont présentés dans la table 1. Nous les rapportons au nombre de tokens traités puisqu'en situation réelle les coûts de mise au point sont constants vis-à-vis du nombre de documents traités en production :

Corpus	Tokens	Phase	CO2e (g)	Energie (Wh)	Energie (Wh par token)
Mise au point	52025 (1776 fournis + 50249 générés)	Augmentation du corpus (ou annotation LLM)	115,01	2240	**0,04**
		Grid-search	143,24	2790	0,05
		Fine-tuning	5,12	99	0,001
		Sous-total	263,38	5129	0,1
Production	5946	Inférences	0,244	4,77	**0,0008**

TABLE 1 : Coûts pour chaque phase de traitement

On voit notamment que le ratio entre les coûts d'inférence du système de production (0,0008 Wh/t) et ceux qui sont nécessaires à une annotation équivalente via les modèles génératifs (0,4 Wh/t) est très important (de l'ordre de 1/50).

4.2 Évaluation

Les résultats communiqués sur le corpus de test sont les suivants :

	Macro			Micro		
	Précision	**Rappel**	**F1**	**Précision**	**Rappel**	**F1**
Annotation LLM (run 1)	33,33	25,49	**28,24**	26,12	14,80	18,89
Annotation BERT (run 2)	24,22	20,07	21,38	29,64	17,59	**22,08**

TABLE 2 : Résultats d'évaluation sur le corpus de test

4.3 Discussion

Les performances obtenues sont considérablement inférieures à ce que nous mesurons habituellement sur nos cas métiers, qui se basent généralement sur une typologie d'entités moins ambitieuse mais dans un contexte miltilingue.

Dans le cadre de ce challenge, ces résultats sont toutefois peu surprenants compte tenu de la complexité induite par le modèle d'annotation cible (cf. section 2), et du fait que les instructions (*prompts*) que nous avons utilisées pour la partie réalisée via des LLM n'incorporent pas toutes les connaissances du domaine qui auraient probablement été utiles pour lever les ambiguïtés entre certaines étiquettes.

En revanche il est intéressant de constater que les performances de la méthode de production sont analogues à celles obtenues en appliquant la méthode d'annotation par LLM que nous utilisons normalement pendant la phase de mise au point. C'est un élément qui tend à confirmer la pertinence de l'approche en termes d'économies de ressources, à performances égales.

Dans la continuité de ces travaux il serait intéressant de reproduire cette même comparaison sur la base de prompts plus élaborés, incorporant notamment les critères donnés dans le guide d'annotation. Nous espérons que cela permettrait de faire évoluer les performances du modèle de mise au point, et celles du modèle de production dans les mêmes proportions.

Un autre point intéressant concerne l'inversion des performances entre les deux approches selon que l'on considère macro-F1 ou micro-F1. Cela s'explique par le fait que pour l'approche BERT, l'ajustement sur les nouvelles classes dépend fortement de la représentation de celles-ci, or nous avons observé un fort déséquilibre dans leur distribution, et que pour certaines entités le modèle avait un score proche de zéro en raison de ce support trop faible. L'approche LLM n'a évidemment pas cette sensibilité au déséquilibre de classes.

5 Conclusion

Face à ce challenge qui combine des contraintes fortes en termes de taille de corpus aussi bien que de finesse du modèle d'annotation, ce qui oriente naturellement vers les grands modèles de langage, nous avons expérimenté une approche hybride qui présente l'intérêt d'utiliser le paradigme génératif avec parcimonie et ainsi de rendre l'ensemble économiquement viable dans un contexte de traitement de données massives et sensibles.

La comparaison entre ce système et une approche basée uniquement sur des LLMs nous conforte dans sa validité de principe, dans le sens où elle offre une économie de ressources considérable sans pour autant dégrader significativement la qualité obtenue en première instance, a fortiori si on considère le micro-F1.

Les scores obtenus sont toutefois assez faibles dans l'absolu, et en tout état de cause largement inférieurs à ce que nous observons généralement sur nos cas d'utilisation réels. Ce résultat nous incite à poursuivre nos investigations sur la conception des prompts afin de prendre en considération les difficultés rencontrées devant une typologie d'entités dont les frontières sont plus fines que ce que nous rencontrons habituellement.

Références

Akbik, A., Bergmann, T., Blythe, D., Rasul, K., Schweter, S. & Vollgraf, R. (2019). FLAIR: An easy-to-use framework for state-of-the-art NLP. *Preprint ACLAnthology:N19-4010*.

Xu, D., Chen, W., Peng, W., Zhang, C., Xu, T., Zhao, X. & Chen, E. (2023). Large language models for generative information extraction: A survey. *Preprint arXiv:2312.17617*

Yahiaoui F. et Limouni E. (2024). Génération et annotation de corpus pour l'entraînement et l'évaluation de modèles d'extraction de relations : utilisation de bibliothèques de génération de données et de LLMs. *Soumis à l'atelier EvalLLM 2024.*

Zaratiana, U., Tomeh, N., Holat, P., & Charnois, T. (2023). Gliner: Generalist model for named entity recognition using bidirectional transformer. *Preprint arXiv:2311.08526*.