Offer #2024-08502

Stage en Augmentation de données textuelles à l'aide de Grands Modèles de Langage (LLM) (F/H)

The offer description below is in French

Contract type : Internship

Level of qualifications required : Bachelor's degree or equivalent

Fonction : Internship Research

Context

Inria Défense&Sécurité (Inria D&S) a été créé en 2020 pour fédérer les actions d'Inria répondant aux besoins numériques des forces armées et forces de l'intérieur. Le stage sera réalisée au sein de l'équipe de recherche en TALN de Inria D&S en partenariat avec l'Agence Ministérielle pour l'IA de Défense (AMIAD), sous la direction de Pauline Soutrenon et Lucie Chasseur Ingénieures NLP (Mission D&S Inria, Grenoble) ainsi que Nihel Kooli Experte NLP et IA (Agence Ministérielle pour l'IA de Défense (AMIAD, Rennes)).

Dans le domaine du Traitement Automatique des Langues, le développement de systèmes performants repose fortement sur des jeux de données annotées de haute qualité. Ces annotations, qui peuvent inclure des étiquettes de catégories, des entités nommées ou des relations syntaxiques, sont essentielles pour permettre aux modèles d'apprentissage automatique de capturer les subtilités du langage humain. Cependant, constituer ces jeux de données est une tâche complexe, chronophage et coûteuse, nécessitant une expertise linguistique, une standardisation rigoureuse et des efforts considérables pour garantir la cohérence des annotations. Ces contraintes rendent l'accès à des données de qualité particulièrement difficile, notamment pour les langues peu représentées ou les domaines spécialisés comme celui de la défense.

Cette rareté des données annotées constitue une problématique majeure dans le domaine du Traitement Automatique des Langues. Quelle que soit la tâche, l'efficacité des approches repose sur la disponibilité des données annotées. Dans la plupart des cas, ces données sont limitées ou parfois même inexistantes, ce qui représente un frein important au développement de solutions robustes.

L'émergence des Grands Modèles de Langage (LLM), tels que ChatGPT, Llama ou Mistral, offre une opportunité de générer, enrichir ou diversifier des jeux de données de manière automatisée tout en réduisant les coûts et les délais associés à leur production.

Ce stage s'inscrit dans cette perspective et a pour objectif d'explorer les capacités des LLM pour répondre aux besoins critiques de données annotées.

Assignment

Ce projet s'inscrit dans la continuité de notre participation audéfi TextMine 2025 pour lequel nous avons mené des travaux d'augmentation de données du jeu de données fourni avec un LLM afin d'optimiser les performances d'un modèle d'extraction de relations. Ces travaux ont produit des résultats prometteurs qui nécessitent d'être approfondis.

Dans le cadre de ce stage, l'augmentation de données sera appliquée à un corpus de documents textuels issus d'informations en direct du journal Le Monde. Ces news ont été produites sur la période initiale de la guerre en Ukraine entre février et mars 2022. L'accumulation de données journalistiques depuis le début de la guerre en Ukraine offre l'opportunité de constituer de nouveaux corpus où le vocabulaire spécialisé de la défense est omniprésent. Un premier travail d'annotation du corpus a été effectué en suivant un guide d'annotation spécifiquement conçu pour ce projet.

La première partie du stage consistera à prendre connaissances des données. Des tests pourront ensuite être réalisés pour sélectionner le ou les LLM les plus pertinents pour cette tâche. Puis, la mission consistera à définir une stratégie d'augmentation (prompt(s) à utiliser, processus pour garantir la qualité et la cohérence des données générées) et à mettre en place la pipeline d'augmentation de données (en mettant l'accent sur le traitement et le formatage des réponses générées par le LLM). Enfin, une évaluation de l'impact des données générées pourra être réalisée de manière à vérifier la qualité des annotations et à identifier les biais et les cas limites.

Main activities

- Analyser les besoins et se familiariser avec les données

- Tester et sélectionner le(s) LLM le(s) plus pertinent(s) pour cette tâche
- Définir une stratégie d'augmentation de données
- Mettre en place la pipeline d'augmentation de données
- Évaluer l'impact des données générées
- Documenter et présenter les résultats

Skills

- Maîtrise du français écrit et parlé
- Connaissances solides en linguistique et en traitement automatique des langues
- Connaissance de Python
- Familiarité avec les LLM et leurs API
- Une connaissance d'outil d'annotation, comme Label studio, sera appréciée

Références

Armingaud, R., Peuvot, A., Besançon, R., Ferret, O., Souihi, S., et al. (2024, July). CEA-List@EvalLLM2024 : prompter un très grand modèle de langue ou affiner un plus petit ?. Atelier sur l'évaluation des modèles génératifs (LLM) et challence d'extraction d'information few-shot, Institut des sciences informatiques et de leurs interactions - CNRS Sciences informatiques [INS2I-CNRS], Toulouse, France.

Bogdanov, S., Constantin, A., Bernard, T., Crabb'e, B., & Bernard, E. (2024, February). NuNER: Entity Recognition Encoder Pre-training via LLM-Annotated Data. *In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing.*

Dai, X., & Adel, H. (2020, October). An Analysis of Simple Data Augmentation for Named Entity Recognition. *In Proceedings of the 28th International Conference on Computational Linguistics.*

Ye, J., Xu, N., Wang, Y., Zhou, J., Zhang, Q., Gui, T., & Huang, X. (2024, February). LLM-DA: Data Augmentation via Large Language Models for Few-Shot Named Entity Recognition.

Benefits package

- Restauration subventionnée
- Transports publics remboursés partiellement
- Congés déterminés en fonction de la durée du stage
- Équipements professionnels à disposition (visioconférence, prêts de matériels informatiques, etc.) Prestations sociales, culturelles et sportives (Association de gestion des œuvres sociales d'Inria

General Information

- **Town/city :** Grenoble
- **Inria Center :** Siège
- **Starting date :** 2025-02-01
- **Duration of contract :** 6 months
- **Deadline to apply :** 2025-02-28

Contacts

- **Inria Team :** MIS-DEFENSE (DIRECTION)
- **Recruiter :**
  Arunraja Emilie / emilie.arunraja@inria.fr

About Inria

Inria is the French national research institute dedicated to digital science and technology. It employs 2,600 people. Its 200 agile project teams, generally run jointly with academic partners, include more than 3,500 scientists and engineers working to meet the challenges of digital technology, often at the interface with other disciplines. The Institute also employs numerous talents in over forty different professions. 900 research support staff contribute to the preparation and development of scientific and entrepreneurial projects that have a worldwide impact.

Warning : you must enter your e-mail address in order to save your application to Inria. Applications must be submitted online on the Inria website. Processing of applications sent from other channels is not guaranteed.

Instruction to apply

Defence Security :
This position is likely to be situated in a restricted area (ZRR), as defined in Decree No. 2011-1425 relating

to the protection of national scientific and technical potential (PPST).Authorisation to enter an area is granted by the director of the unit, following a favourable Ministerial decision, as defined in the decree of 3 July 2012 relating to the PPST. An unfavourable Ministerial decision in respect of a position situated in a ZRR would result in the cancellation of the appointment.

Recruitment Policy :
As part of its diversity policy, all Inria positions are accessible to people with disabilities.