Mickaël BARON
École Nationale Supérieure de Mécanique et d'Aérotechnique
LIAS – EA 6315
1 avenue Clément Ader BP 40109
86961 Chasseneuil Cedex – France
Tél : +33 (0)5 49 49 80 73
baron@ensma.fr

Construction et exploitation de dictionnaires de composants au travers de l'IA générative

Localisation : Site du Futuroscope, France
Date de début : octobre 2025 (flexible)
Encadrants : Stéphane JEAN et Mickaël BARON *(l'encadrant dont le nom est souligné est l'encadrant référent : baron@ensma.fr)*
Mots clés : IA générative, ontologie

Contexte du stage

Le CRITT Informatique (Centre Régional d'Innovation et de Transfert de Technologie) est une structure dédiée au transfert de technologie, labellisée CRT (Centre de Ressources Technologiques) au niveau national par le ministère de la Recherche. Ces centres ont été créés pour accompagner les entreprises dans leur développement et leur transformation technologique.

Dans ce contexte, le CRITT Informatique est régulièrement sollicité par des entreprises du domaine industrielle (aéronautique, ferroviaire, militaire, etc.) pour les accompagner dans la création de dictionnaires de composants, facilitant ainsi les échanges entre les entreprises. Ces dictionnaires de composants appelés des ontologies peuvent s'appuyer sur des standards comme par exemple OntoML (ISO 13584-32). Un travail de mapping est réalisé pour définir les ontologies à partir des concepts et des données de l'entreprise en accord avec ces standards. Ce travail est coûteux en temps, car la compréhension des données métiers est souvent complexe et nécessite alors l'analyse de nombreux documents structurés et non structurés lorsqu'ils sont disponibles.

L'utilisation des IA génératives de type LLM (Large Language Model) pourrait assister les équipes du CRITT dans la construction des ontologies [1]. C'est pourquoi le CRITT a sollicité le LIAS (Laboratoire Informatique et d'Automatique pour les Systèmes), qui travaille déjà sur l'application des LLM aux données historiques et sur le mapping de ces données pour construire des ontologies. Plus spécifiquement, les travaux du LIAS dans le cadre du projet ANR Digitalis [1] s'appuient sur un modèle conceptuel répandu CDOC-CRM ainsi que sur des IA génératives commerciales, comme GPT. L'utilisation de ces IA générative facilite le travail, car elles possèdent déjà une connaissance du modèle CDOC-CRM.

Objectifs de la thèse

Un des enjeux principaux de cette thèse est de considérer que l'IA générative ne possède pas de connaissance préalable du modèle qui définit la structure de l'ontologie [2]. Il est possible d'explorer l'extension des connaissances de ces IA génératives en s'appuyant, par exemple, sur le contenu du standard OntoML. Ainsi, un objectif à envisager serait d'utiliser la technique RAG

1. https ://digitalis.humanities.science/

Mickaël BARON
École Nationale Supérieure de Mécanique et d'Aérotechnique
LIAS – EA 6315
1 avenue Clément Ader BP 40109
86961 Chasseneuil Cedex – France
Tél : +33 (0)5 49 49 80 73
baron@ensma.fr

(Retrieval Augmented Generation) pour alimenter en contexte la phase d'interrogation de l'IA générative. Toutefois, cette solution se heurte souvent à des problèmes de contextes insuffisants ou trop importants pour que l'IA générative puisse répondre efficacement. Ces problèmes de contexte ont été abordés au LIAS dans un autre domaine [3] : les bases de connaissances sémantiques. Cette piste de recherche permettrait d'affiner le contexte à transmettre à l'IA générative et d'expliquer les raisons d'un manque ou d'un excès de contexte.

Dans un souci de confidentialité des données traitées par les IA génératives, souvent exigée par les entreprises, l'inférence sur site (on-premise) des IA génératives représente un autre défi. Le défi du déploiement et de la qualité des résultats des modèles d'IA générative inférés est d'assurer une performance équivalente à celle des solutions commerciales. Par ailleurs, l'inférence sur site permettra d'étudier l'efficacité énergétique des solutions mise en oeuvre, car le CRITT Informatique doit répondre aux exigences régionales en la matière.

Le cadre de cette thèse Cifre regroupe des problématiques de recherche liées à la construction d'ontologies via des IA génératives. Les besoins du CRITT Informatique incluent également le développement d'outils informatiques pour faciliter la mise en oeuvre des recherches obtenues dans cette thèse.

Profil du candidat et documents à fournir

Le candidat devra être titulaire d'un Master 2 ou d'un diplôme d'ingénieur et posséder des connaissances en développement logiciel et manipulation de modèles d'IA générative. Un bon niveau en français et en anglais est nécessaire. Les documents à fournir :

- Curriculum Vitae ;
- Notes de Master ou équivalent avec si possible le classement ;
- Tout document pertinent pour compléter la candidature (ex. rapports, dépôt GitHub).

Le dossier de candidature doit être envoyé avant le 2 juin à 9h par email à baron@ensma.fr

Références

[1] X. Liang, Z. Wang, M. Li, and Z. Yan, "A survey of LLM-augmented knowledge graph construction and application in complex product design", *Procedia CIRP*, vol. 128, pp. 870–875, 2024, 34th CIRP Design Conference, ISSN: 2212-8271. DOI: https://doi.org/10.1016/j.procir.2024.07.069. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2212827124007911.

[2] V. K. Kommineni, B. König-Ries, and S. Samuel, "From human experts to machines: an LLM supported approach to ontology and knowledge graph construction", *CoRR*, vol. abs/2403-.08345, 2024. DOI: 10.48550/ARXIV.2403.08345. arXiv: 2403.08345. [Online]. Available: https://doi.org/10.48550/arXiv.2403.08345.

[3] L. Parkin, "Techniquess coopératives pour l'exploitation des bases de connaissances et passage à l'échelle", Theses, ISAE-ENSMA Ecole Nationale Supérieure de Mécanique et d'Aérotechique - Poitiers, Dec. 2022. [Online]. Available: https://theses.hal.science/tel-03934427.