Using Large Language Models for Few-shot Entity and Relation Extraction in Scientific Articles on New Material Design

6-month internship @ CEA List

Internship context

Based in Saclay (Essonne), the LIST is one of the two institutes of CEA Tech, the Technological Research Division of the CEA. Dedicated to intelligent digital systems, its mission is to carry out technological developments of excellence on behalf of industrial partners, in order to create value.

Within the LIST, the Laboratory of Textual and Visual Semantic Analysis (LASTI) conducts its research in the field of natural language processing and computer vision to extract, classify and generate information. The laboratory's research themes include learning with few data, trustworthiness and multimodality.

Missions

In the domain of materials science, designing innovative materials very often starts with a study of the state of the art, in order to find out the positioning of existing materials, more precisely their composition, properties or manufacturing parameters. The goal of this internship is to explore how Artificial Intelligence and Natural Language Processing can help in automating this tedious task, by identifying relevant entities (e.g., materials, chemical compounds, synthesis methods) and their relationships (e.g., material properties, synthesis outcomes) in the unstructured scientific texts.

The selected intern will work on the following tasks :

- Explore the state-of-the-art of LLMs (specific encoder models such as BERT-based models or large instructed models) for entity and relation extraction from scientific articles, with a focus on few-shot learning methods.
- Fine-tune pre-trained LLMs for few-shot tasks, ensuring accurate identification of scientific terms and their relationships within the domain of material science.
- Design and implement pipelines for automatic extraction, validation, and integration of scientific knowledge from unstructured text into structured formats (e.g., knowledge graphs, databases).
- Explore existing datasets from scientific publications related to the design of new materials, focusing on extracting relevant entities and relationships with minimal labeled data and evaluate the developed approaches on these datasets
- Work in collaboration with domain experts in materials science to ensure the relevance and utility of the extracted information.

Qualifications

- Students in their 4th or 5th year of studies (M1, M2 or gap year)
- Knowledge in Natural Language Processing
- Machine learning skills (deep learning, perception models, generative AI…)
- Python proficiency in a deep learning framework (especially PyTorch or Tensorflow)

Job-related benefits

Joining the CEA List and the LVA as an intern means:

- Working in one of the most innovative research organizations in the world, addressing societal challenges to build the world of tomorrow
- Discovering a rich ecosystem: privileged connections between the industrial and academic sectors
- Conducting research autonomously and creatively: encouragement to valorize results (scientific articles, patents, open-source codes...)
- Join a young and dynamic team
- Benefit from an internal computing infrastructure with more than 300 state-of-the-art GPUs
- Receive a stipend between €1300 and €1400 per month
- Have the opportunity to continue with a PhD or as a research engineer after the internship
- Work remotely (2 days/week)
- Receive a 75% reimbursement on public transportation costs, and benefit from the “mobili-jeune” aid to reduce rent costs…

To apply, contact the laboratory with a CV and cover letter: romaric.besancon@cea.fr

Information extraction from complex documents

6-month internship @ CEA List

Internship context

Based in Saclay (Essonne), the LIST is one of the two institutes of CEA Tech, the Technological Research Division of the CEA. Dedicated to intelligent digital systems, its mission is to carry out technological developments of excellence on behalf of industrial partners, in order to create value.

Within the LIST, the Laboratory of Textual and Visual Semantic Analysis (LASTI) conducts its research in the field of natural language processing and computer vision to extract, classify and generate information. The laboratory's research themes include learning with few data, trustworthiness and multi-modality using discriminative and generative AI.

Topic and objectives

Portable Document Format (PDF) files have become the standard for sharing documents. Their popularity is based on their ability to preserve formatting, which ensures that the content appears the same regardless of where it is viewed. Nevertheless this feature makes parsing and extracting information from PDFs challenging.
The purpose of this internship is to study the principles, approaches and methods of information extraction from PDFs (e.g combining traditional Optical Character Recognition (OCR) and extraction information, or using large language models or vision models).

The intern will select a few promising methods, and evaluate them, by identifying relevant entities (e.g., materials, compounds, locations) and their relationships in unstructured texts and semi-structured text (such as tables) from PDFs.

After an initial step dedicated to bibliographic research, the internship objectives will be to:

- Prepare public and/or industrial datasets for evaluation and/or training for extraction of specific terms and their relationships.
- Design and implement pipelines for automatic extraction, validation and comparison of methods
- Evaluate the developed approaches on the identified datasets

Qualifications

- M2 or final year of engineering degree
- Knowledge in Natural Language Processing
- Machine learning skills (deep learning, multimedia models, generative AI…)
- Python proficiency in a deep learning framework (especially PyTorch or Tensorflow)

Job-related benefits

Joining the CEA List and the LASTI means:

- Working in one of the most innovative research organizations in the world, addressing societal challenges to build the world of tomorrow
- Discovering a rich ecosystem: privileged connections between industry and academia
- Conducting research autonomously and creatively: encouragement to promote results (scientific articles, patents, open-source codes...)
- Work in a young and dynamic team
- Benefit from an internal computing infrastructure with more than 300 state-of-the-art GPUs
- Receive a monthly stipend of ~1300€
- Have the opportunity to continue with a PhD or as a research engineer after the internship
- Work remotely (2 days/week)
- Receive a 75% reimbursement on public transportation, and benefit from the “mobili-jeune” aid to reduce rent

To apply, send us a CV and cover letter: Anne-Laure.DAQUO@cea.fr , Benjamin.LABBE@cea.fr

Spatio-temporally controlled visual content generation

6-month internship @ CEA List

Internship context

Based in Saclay (Essonne), the LIST is one of the two institutes of CEA Tech, the Technological Research Division of the CEA. Dedicated to intelligent digital systems, its mission is to carry out technological developments of excellence on behalf of industrial partners, in order to create value.

Within the LIST, the Laboratory of Textual and Visual Semantic Analysis (LASTI) conducts its research in the field of natural language processing and computer vision to extract, classify and generate information. The laboratory's research themes include learning with few data, trustworthiness and multimodality using discriminative and generative AI.

Topic and objectives

The visual representation of concepts (e.g objects) varies in space and time. For instance, new car models appear, and the same brands and models are not popular in different regions of the world. These variations are not well addressed by the current generative AI models. They are also under-researched in computer vision in general. The objective of the internship is to explore the spatio-temporal grounding of visual concepts and to propose improved generation methods along these dimensions. Scientifically, this problem relates to lifelong-learning and fine-grained classification, two areas in which the LASTI lab researchers have proven expertise. The internship will build on this expertise to propose an innovative approach of the topic.

After an initial step dedicated to bibliographic research, the internship objectives will be to:

- Analyze and improve an existing spatially and temporally diversified visual dataset
- Model the spatio-temporal variability of visual concepts in an actionable way
- Propose methods that improve the spatio-temporal grounding of synthetic content by adapting pretrained visual generators
- Introduce dedicated protocols for the comprehensive and sound evaluation of the contribution

The internship is designed as an introduction to research, with the aim of publishing a scientific article if the results are conclusive. This initial work could be continued in a PhD thesis, within the broader framework of AI-related topics addressed by the LASTI lab.

Qualifications

- M2 or final year of engineering degree
- Machine learning skills (deep learning, multimedia models, generative AI…)
- Preferred: hands-on experience with visual content generation models
- Python proficiency in a deep learning framework (preferably PyTorch)

Job-related benefits

Joining the CEA List and the LASTI means:

- Working in one of the most innovative research organizations in the world, addressing societal challenges to build the world of tomorrow
- Discovering a rich ecosystem: privileged connections between industry and academia
- Conducting research autonomously and creatively: encouragement to promote results (scientific articles, patents, open-source codes...)
- Work in a young and dynamic team
- Benefit from an internal computing infrastructure with more than 300 state-of-the-art GPUs
- Receive a monthly stipend of ~1300€
- Have the opportunity to continue with a PhD or as a research engineer after the internship
- Work remotely (2 days/week)
- Receive a 75% reimbursement on public transportation, and benefit from the “mobili-jeune” aid to reduce rent

To apply, send us a CV and cover letter: herve.le-borgne@cea.fr and adrian.popescu@cea.fr

Few-shot Named Entity Recognition

6-month internship @ CEA List

Internship context

Based in Saclay (Essonne), the LIST is one of the two institutes of CEA Tech, the Technological Research Division of the CEA. Dedicated to intelligent digital systems, its mission is to carry out technological developments of excellence on behalf of industrial partners, in order to create value.

Within the LIST, the Laboratory of Textual and Visual Semantic Analysis (LASTI) conducts its research in the field of natural language processing and computer vision to extract, classify and generate information. The laboratory's research themes include learning with few data, trustworthiness and multimodality using discriminative and generative AI.

Missions

The results of the EVALLLM 2024 challenge, in which the LASTI team played an active role, have shown that encoder-based models such as BERT outperform large decoder-based models, particularly in few-shot learning scenarios. These findings pave the way for the exploration of new encoder-based architectures to improve Named Entity Recognition (NER).

Based on the GLiNER model, which was successfully used in EVALLLM 2024, the internship aims at investigating various encoder-based model architectures in order to improve few-shot NER performance. Therefore, the intern will work on the following main tasks:

- Conduct a literature review on encoder-based models (e.g., BERT) and few-shot approaches in NER;
- Perform an in-depth analysis of the GLiNER model used during EVALLLM 2024 to identify potential improvements;
- Design and develop new encoder-based architectures for few-shot NER;
- Conduct experiments and performance evaluations on standard datasets as well as on those used in EVALLLM 2024;
- Compare results against large decoder-based models, providing insights and recommendations for further developments.

This internship may be seen as an introduction to research. It may lead to the publishing a scientific paper if the results are convincing. Successful outcomes may also lead to the extension of this work through a PhD thesis, within the broader scope of the LASTI team's AI research.

Qualifications

- M2 or final year of engineering degree
- Machine learning skills (deep learning, generative AI…)
- Knowledge in Natural Language Processing
- Python proficiency in a deep learning framework (especially PyTorch or Tensorflow)

Job-related benefits

Joining the CEA List and the LASTI means:

- Working in one of the most innovative research organizations in the world, addressing societal challenges to build the world of tomorrow
- Discovering a rich ecosystem: privileged connections between industry and academia
- Conducting research autonomously and creatively: encouragement to promote results (scientific articles, patents, open-source codes...)
- Work in a young and dynamic team
- Benefit from an internal computing infrastructure with more than 300 state-of-the-art GPUs
- Receive a monthly stipend of ~1300€
- Have the opportunity to continue with a PhD or as a research engineer after the internship
- Work remotely (2 days/week)
- Receive a 75% reimbursement on public transportation, and benefit from the “mobili-jeune” aid to reduce rent

To apply, send us a CV and cover letter: olivier.ferret@cea.fr and sondes.souihi@cea.fr

Injecting Symbolic Knowledge into Zero-Shot and Few-Shot Large Language Models

6-month internship @ CEA List

Internship context

Based in Saclay (Essonne), the LIST is one of the two institutes of CEA Tech, the Technological Research Division of the CEA. Dedicated to intelligent digital systems, its mission is to carry out technological developments of excellence on behalf of industrial partners, in order to create value.

Within the LIST, the Laboratory of Textual and Visual Semantic Analysis (LASTI) conducts its research in the field of natural language processing and computer vision to extract, classify and generate information. The laboratory's research themes include learning with few data, trustworthiness and multimodality using discriminative and generative AI.

Missions

Large-scale language models (LLMs) such as GPT and BERT have demonstrated impressive capabilities in a variety of automatic natural language processing (NLP) tasks. However, these models operate mainly on implicit knowledge bases learned from large amounts of unstructured data. Although these models perform well in many situations, their lack of explicit understanding and inability to reason logically pose limitations. The injection of symbolic knowledge (logical rules, knowledge graphs or ontologies...) into LLMs could potentially fill these gaps, improving their ability to reason and understand more robustly, even in zero-shot (no examples) or few-shot (few examples) scenarios.

The aim of this internship is to explore methods for efficiently integrating symbolic knowledge into large language models, particularly in zero-shot and few-shot contexts. The aim is to enable LLMs to exploit structured and explicit knowledge to improve their performance on specific tasks, where training data is scarce or absent.

Work to be carried out will include a literature review on methods for injecting symbolic knowledge into LLMs; exploration of zero-shot and few-shot approaches in the context of LLMs; development of a method for integrating symbolic knowledge into LLMs; implementation of knowledge injection algorithms in zero-shot and few-shot scenarios; comparing LLM performance with and without the integration of symbolic knowledge on specific tasks; analyzing the benefits of symbolic knowledge in terms of accuracy, robustness and generalizability; proposing improvements to knowledge injection techniques, taking into account the limitations identified during the experiments.

The experiments will be carried out on our FactoryAI cluster, comprising several dozen multi-GPU nodes.

Qualifications

- M2 or final year of engineering degree
- Machine learning skills (deep learning, generative AI…)
- Knowledge in Natural Language Processing
- Python proficiency in a deep learning framework (especially PyTorch or Tensorflow)

Job-related benefits

Joining the CEA List and the LASTI means:

- Working in one of the most innovative research organizations in the world, addressing societal challenges to build the world of tomorrow
- Discovering a rich ecosystem: privileged connections between industry and academia
- Conducting research autonomously and creatively: encouragement to promote results (scientific articles, patents, open-source codes...)
- Work in a young and dynamic team
- Benefit from an internal computing infrastructure with more than 300 state-of-the-art GPUs
- Receive a monthly stipend of ~1300€
- Have the opportunity to continue with a PhD or as a research engineer after the internship
- Work remotely (2 days/week)
- Receive a 75% reimbursement on public transportation, and benefit from the “mobili-jeune” aid to reduce rent

To apply, send us a CV and cover letter: gael.de-chalendar@cea.fr **and** evan.dufraisse@cea.fr

Frugal learning and adaptation of textual generative models

6-month internship @ CEA List

Internship context

Based in Saclay (Essonne), the LIST is one of the two institutes of CEA Tech, the Technological Research Division of the CEA. Dedicated to intelligent digital systems, its mission is to carry out technological developments of excellence on behalf of industrial partners, in order to create value.

Within the LIST, the Laboratory of Textual and Visual Semantic Analysis (LASTI) conducts its research in the field of natural language processing and computer vision to extract, classify and generate information. The laboratory's research themes include learning with few data, trustworthiness and multimodality using discriminative and generative AI.

Topic and objectives

Generative language models have long been at the center of a competition focused on the number of parameters, considered central to improving generalization and performance [1](Kaplan et al. 2020). However, DeepMind's new scaling laws [2](Hoffman et al. 2022), at the origin of the Chinchilla model, have relativized this factor by optimizing the triplet (model size, amount of data, and computational resources). However, these studies overlook the impact of pre-training data quality. Other research shows that it is possible to reduce pre-training data by selecting high-quality data without loss of performance [3](Marion et al. 2023), [4](Sorscher et al. 2022), confirming the validity of fixed-data scaling laws.

These preliminary results suggest that data optimization could reduce the resources required for model convergence, a major constraint due to the cost of training. The internship therefore focuses on the development of resource-efficient fine-tuning and pre-training methods based on data optimization. The candidate will explore the techniques of Data-Pruning [3](Marion et al. 2023), Curriculum-Learning [6](Wang et al. 2021), and model distillation [7](Gou et al. 2021).

During its internship the trainee will :

- Familiarize himself/herself with these techniques through a literature review.
- Design a training strategy in collaboration with its supervisors.
- Test this strategy on the CEA's FactoryIA cluster.
- Iteratively improve the strategy based on the results.

Qualifications

Engineering degree and/or M2 in computer science with a strong interest in artificial learning.
Skills required :
- working environment: linux
- basic notions of machine learning and neural networks.
- programming: Python + PyTorch

Job-related benefits

Joining the CEA List and the LASTI means:

- Working in one of the most innovative research organizations in the world, addressing societal challenges to build the world of tomorrow
- Discovering a rich ecosystem: privileged connections between industry and academia
- Conducting research autonomously and creatively: encouragement to promote results (scientific articles, patents, open-source codes...)
- Work in a young and dynamic team
- Benefit from an internal computing infrastructure with more than 300 state-of-the-art GPUs
- Receive a monthly stipend of ~1300€
- Have the opportunity to continue with a PhD or as a research engineer after the internship
- Work remotely (2 days/week), receive a 75% reimbursement on public transportation, and benefit from the “mobili-jeune” aid to reduce rent

To apply, send us a CV and cover letter: evan.dufraisse@cea.fr

Toward robust fact checking in domain-specific news

6-month internship @ CEA List

Internship context

Based in Saclay (Essonne), the LIST is one of the two institutes of CEA Tech, the Technological Research Division of the CEA. Dedicated to intelligent digital systems, its mission is to carry out technological developments of excellence on behalf of industrial partners, in order to create value.

Within CEA LIST, the LASTI lab conducts its research in the field of natural language processing and computer vision to extract, classify and generate information. The LISA lab conducts research in the field of haptic and audio interaction and user experience.

Topic and objectives

Citizens' use of online media to get informed comes with opportunities in terms of diversity, but also challenges regarding the accuracy of the consumed news. Disinformation is one such risk, and the availability of large language models (LLMs) facilitates the set-up of coordinated campaigns via the generation of linguistically plausible texts. Existing research focuses on detecting the truthfulness or falseness of news. While interesting, this approach is insufficient for texts discussing complex topics or events. Such news build arguments based on multiple affirmations whose validity and factuality should be checked to establish whether there is a disinformation intent. Given the massive amount of online news, manual verification is only possible for a small subset of documents, and the automation of the process remains challenging. The internship objectives will be to:

- Analyze citizens' attitudes towards false news
- Analyze existing manual fact checking and automatic approaches, focusing on complex topics and events and taking into account aspects such as truthfulness, knowledge obsolescence, author intent, source reliability, etc.
- Propose an LLM-based method for fact-checking automation in a specific domain, such as health-related news, that provides incorporates at least one of above aspects and provides detailed explanations about the reasoning used to analyze the news
- Study the efficacy of the proposed method quantitatively, using annotated fact-checking datasets, and qualitatively, via user-centered experiments
- Showcase the results through a prototype

The internship is designed as an introduction to research, with the aim of publishing a scientific article if the results are conclusive. This initial work could be continued in a PhD thesis, within the broader framework of AI-related topics addressed by the LASTI or LISA lab.

Qualifications

- M2 or final year of engineering degree
- Machine learning skills (deep learning, LLMs, generative AI…)
- Preferred: hands-on experience with natural language processing and interest in human-computer interaction
- Python proficiency in a deep learning framework (preferably PyTorch)

Job-related benefits

Joining the CEA List and the LASTI means:

- Working in one of the most innovative research organizations in the world, addressing societal challenges to build the world of tomorrow
- Discovering a rich ecosystem: privileged connections between industry and academia
- Conducting research autonomously and creatively: encouragement to promote results (scientific articles, patents, open-source codes...)
- Work in a young and dynamic team
- Benefit from an internal computing infrastructure with more than 300 state-of-the-art GPUs
- Receive a monthly stipend of ~1300€
- Have the opportunity to continue with a PhD or as a research engineer after the internship
- Work remotely (2 days/week),
- receive a 75% reimbursement on public transportation, and benefit from the "mobili-jeune" aid to reduce rent

To apply, send us a CV and cover letter: adrian.popescu@cea.fr and sabrina.paneels@cea.fr

Apprentissage et adaptation frugales de modèles génératifs textuels

Stage de 6 mois @ CEA List

Contexte

Basé à Saclay (Essonne), le LIST est l'un des deux instituts de CEA Tech, la Direction de la Recherche Technologique du CEA. Dédié aux systèmes numériques intelligents, il a pour mission de réaliser des développements technologiques d'excellence pour le compte de partenaires industriels, afin de créer de la valeur. Au sein du LIST, le Laboratoire d'analyse sémantique textuelle et visuelle (LASTI) mène ses recherches dans le domaine du traitement du langage naturel et de la vision par ordinateur pour extraire, classer et produire de l'information. Les thèmes de recherche du laboratoire comprennent l'apprentissage avec peu de données, la fiabilité et la multimodalité à l'aide de l'IA discriminative et générative.

Présentation du sujet

Les modèles de langage génératifs ont longtemps été au centre d'une compétition axée sur le nombre de paramètres, considéré comme central pour améliorer la généralisation et la performance [1](Kaplan et al. 2020). Toutefois, les nouvelles lois d'échelle de DeepMind [2](Hoffman et al. 2022), à l'origine du modèle Chinchilla, ont relativisé ce facteur en optimisant le triplet (taille du modèle, quantité de données, et ressources computationnelles). Cependant, ces études négligent l'impact de la qualité des données de pré-entraînement. D'autres recherches montrent qu'il est possible de réduire les données de pré-entraînement en sélectionnant des données de haute qualité sans perte de performance [3](Marion et al. 2023), [4](Sorscher et al. 2022), confirmant la validité des lois d'échelle à données fixes.

Ces résultats préliminaires suggèrent qu'une optimisation des données pourrait réduire les ressources nécessaires à la convergence des modèles, une contrainte majeure de part leur coût d'entrainement. Le stage porte donc sur le développement de méthodes de fine-tuning et de pré-entraînement frugaux en ressources, axées sur l'optimisation des données. Le candidat explorera les techniques de Data-Pruning [3](Marion et al. 2023), de Curriculum-Learning [6](Wang et al. 2021), et de distillation de modèles [7](Gou et al. 2021).
Le stagiaire devra :

- Se familiariser avec ces techniques via une étude bibliographique.
- Concevoir une stratégie d'entraînement en collaboration avec l'équipe.
- Tester cette stratégie sur le cluster FactoryIA du CEA.
- Améliorer itérativement la stratégie en fonction des résultats.

Qualifications

Formation d'ingénieur et/ou M2 en informatique avec un fort intérêt pour l'apprentissage artificiel.
Compétences requises :
• environnement de travail : Linux
• notions de base en apprentissage automatique et en réseaux de neurones.
• programmation : Python + PyTorch

Avantages

Rejoindre la liste du CEA et le LASTI, c'est.. :

- Travailler dans l'un des organismes de recherche les plus innovants au monde, relever des défis sociétaux pour construire le monde de demain.
- Découvrir un écosystème riche : des liens privilégiés entre l'industrie et le monde universitaire
- Mener des recherches de manière autonome et créative : encouragement à promouvoir les résultats (articles scientifiques, brevets, codes open-source...)
- Travailler au sein d'une équipe jeune et dynamique
- Bénéficier d'une infrastructure informatique interne avec plus de 300 GPU de pointe
- Recevoir une allocation mensuelle d'environ 1300 €.
- Avoir la possibilité de poursuivre avec un doctorat ou en tant qu'ingénieur de recherche après le stage
- Travailler à distance (2 jours/semaine), recevoir un remboursement de 75% sur les transports en commun, et bénéficier de l'aide « mobili-jeune » pour réduire le loyer.

Pour postuler, envoyez-nous un CV et une lettre de motivation : evan.dufraisse@cea.fr

Injection de Connaissances Symboliques dans les Modèles de Langage de GrandeTaille en Zero-Shot et Few-Shot

Stage de 6 mois @ CEA List

Contexte du stage

Basé à Saclay (Essonne), le LIST est l'un des deux instituts de CEA Tech, la Direction de la Recherche Technologique du CEA. Dédié aux systèmes numériques intelligents, il a pour mission de réaliser des développements technologiques d'excellence pour le compte de partenaires industriels, afin de créer de la valeur.
Au sein du LIST, le Laboratoire d'analyse sémantique texte et image (LASTI) mène ses recherches dans le domaine du traitement du langage naturel et de la vision par ordinateur pour extraire, classer et produire de l'information. Les thèmes de recherche du laboratoire comprennent l'apprentissage avec peu de données, la fiabilité et la multimodalité à l'aide de l'IA discriminative et générative.

Sujet et objectifs

Les modèles de langage de grande taille (LLM) tels que GPT et BERT ont démontré des capacités impressionnantes dans diverses tâches de traitement automatique du langage naturel. Cependant, ces modèles fonctionnent principalement sur des bases de connaissances implicites apprises à partir de grandes quantités de données non structurées. Bien que ces modèles soient performants dans de nombreuses situations, leur manque de compréhension explicite et leur incapacité à raisonner de manière logique posent des limites. L'injection de connaissances symboliques (règles logiques, graphes de connaissances ou ontologies…) dans les LLM pourrait potentiellement combler ces lacunes en améliorant leur capacité à raisonner et à comprendre de manière plus robuste, même dans des scénarios de zero-shot (sans exemple) ou few-shot (avec peu d'exemples).
L'objectif de ce stage est d'explorer des méthodes pour intégrer efficacement des connaissances symboliques dans des modèles de langage de grande taille, en particulier dans des contextes de zero-shot et few-shot. Le but est de permettre aux LLM d'exploiter des connaissances structurées et explicites pour améliorer leur performance sur des tâches spécifiques, où les données d'entraînement sont rares ou absentes.
Les travaux à réaliser incluront une revue de la littérature sur les méthodes d'injection de connaissances symboliques dans les LLM ; l'exploration des approches zero-shot et few-shot dans le cadre des LLM ; le développement d'une méthode pour l'intégration de connaissances symboliques dans les LLM ; la mise en œuvre d'algorithmes d'injection de connaissances dans des scénarios de zero-shot et few-shot ; la comparaison des performances des LLM avec et sans l'intégration de connaissances symboliques sur des tâches spécifiques ; l'analyse des bénéfices apportés par les connaissances symboliques en termes de précision, de robustesse et de capacité de généralisation ; la proposition d'améliorations des techniques d'injection de connaissances, en tenant compte des limites identifiées lors des expérimentations.
Les expérimentations seront réalisées sur notre cluster FactoryAI, composé de plusieurs dizaines de nœuds multi-GPU.

Qualifications

- Étudiant en master 2 ou dernière année d'école d'ingénieur avec une spécialisation en Intelligence Artificielle, Apprentissage Automatique ou Traitement Automatique du Langage Naturel.
- Connaissances solides en apprentissage profond, en modèles de langage et en représentations symboliques.
- Bonnes compétences en programmation (Python, PyTorch/TensorFlow).
- Intérêt pour la recherche appliquée et la résolution de problèmes complexes.

Avantages liés à l'emploi

Rejoindre le CEA List et le LASTI, c'est :

- Travailler dans l'un des organismes de recherche les plus innovants au monde, relever des défis sociétaux pour construire le monde de demain
- Découvrir un écosystème riche : des liens privilégiés entre l'industrie et le monde universitaire
- Mener des recherches de manière autonome et créative : encouragement à promouvoir les résultats (articles scientifiques, brevets, codes open-source...)
- Travailler au sein d'une équipe dynamique
- Bénéficier d'une infrastructure informatique interne avec plus de 300 GPU de pointe
- Recevoir une allocation mensuelle d'environ 1300 €.
- Avoir la possibilité de poursuivre avec un doctorat ou en tant qu'ingénieur de recherche après le stage
- Travailler à distance (2 jours/semaine), recevoir un remboursement de 75% sur les transports en commun, et bénéficier de l'aide « mobili-jeune » pour réduire le loyer.

Pour postuler, envoyez CV et lettre de motivation : gael.de-chalendar@cea.fr **et** evan.dufraisse@cea.fr

Evaluation de la fiabilité et de la robustesse des grands modèles de langue (LLMs) sur des tâches d'annotation dans le domaine médical

Stage de 6 mois @ CEA List

Contexte du stage

Basé à Saclay (Essonne), le LIST est l'un des deux instituts de CEA Tech, la Direction de la Recherche Technologique du CEA. Dédié aux systèmes numériques intelligents, il a pour mission de réaliser des développements technologiques d'excellence pour le compte de partenaires industriels, afin de créer de la valeur. Au sein du LIST, le Laboratoire d'analyse sémantique texte et image (LASTI) mène ses recherches dans le domaine du traitement du langage naturel et de la vision par ordinateur pour extraire, classer et produire de l'information. Les thèmes de recherche du laboratoire comprennent l'apprentissage avec peu de données, la fiabilité et la multimodalité à l'aide de l'IA discriminative et générative.

Sujet et objectifs

Les modèles de langue tels que BERT et GPT-3 affichent des performances impressionnantes sur une variété de tâches du Traitement Automatique des Langues (TAL) et leur fine-tuning permet de spécialiser ces modèles génériques en modèles performants et spécifiques. Cependant, le fait que le fonctionnement interne de ces modèles est difficile à saisir, constitue un frein à leur utilisation dans des applications nécessitant un niveau élevé de fiabilité et de robustesse comme c'est le cas dans le domaine de la santé. Le stage consistera, d'une part, à constituer un framework pour l'évaluation des performances des grands modèles de langue (LLMs pour Large Language Models) et de leur utilisation dans le domaine de la santé, et d'autre part, à évaluer ces modèles sur des tâches de structuration et d'extraction d'informations à partir de comptes-rendus médicaux. Cette évaluation sera réalisée selon deux approches différentes : une évaluation humaine à petite échelle dans laquelle les prédictions de ces modèles seront comparées à une référence créée manuellement par des praticiens de santé, et une évaluation automatique utilisant des métriques pour les tâches de structuration et d'extraction d'informations.

Le stage se déroulera selon les étapes suivantes:

- Recherche de benchmarks académiques pour l'évaluation des LLMs dans le domaine de la santé.
- Constitution d'un dataset de référence pour l'évaluation des LLMs sur des tâches de structuration et d'extraction d'informations à partir de comptes-rendus médicaux.
- Réalisation d'une évaluation comparative des performances des LLMs en zero ou few shot learning pour le NER sur des comptes rendus médicaux fictifs. Dans l'analyse comparative, des modèles équivalents à GPT 3, GPT 4, falcon, Llama, Bloom peuvent être étudiés.
- Evaluation manuelle (humaine) et automatique des LLMs sur diverses tâches de TAL. Un intérêt particulier sera accordé à l'étude du phénomène d'hallucination et d'omission des LLMs.
- Développer une chaine de traitements pour analyser, interpréter et contrôler les prédictions des LLMs.

Qualifications

- Étudiant en master 2 ou dernière année d'école d'ingénieur
- Maîtrise du langage de programmation Python
- Maîtrise des méthodes d'évaluation des modèles de Machine Learning ou Deep Learning en NLP
- Expérience avec une bibliothèque de type Transfomers, Tensorflow, PyTorch, etc.
- Notions de base en Traitement Automatique des Langues

Avantages liés à l'emploi

Rejoindre le CEA List et le LASTI, c'est:

- Travailler dans l'un des organismes de recherche les plus innovants au monde, relever des défis sociétaux pour construire le monde de demain
- Découvrir un écosystème riche : des liens privilégiés entre l'industrie et le monde universitaire
- Mener des recherches de manière autonome et créative : encouragement à promouvoir les résultats (articles scientifiques, brevets, codes open source...)
- Travailler au sein d'une équipe dynamique
- Bénéficier d'une infrastructure informatique interne avec plus de 300 GPU de pointe
- Recevoir une allocation mensuelle d'environ 1300 €
- Avoir la possibilité de poursuivre avec un doctorat ou en tant qu'ingénieur de recherche après le stage
- Travailler à distance (2 jours/semaine), recevoir un remboursement de 75% sur les transports en commun, et bénéficier de l'aide « mobili-jeune » pour réduire le loyer

Pour postuler, envoyez CV et lettre de motivation: nasredine.semmar@cea.fr