Extraction de variables expérimentales à partir d'un texte scientifique
1 ou 2 étudiants

Constance Douwes, Carlos Ramisch, Alexis Nasr

10 janvier 2026

Le thème général de ce sujet est de concevoir un système permettant de prédire le coût d'expériences de deep learning à partir d'articles scientifiques les décrivant. Etant donné un article scientifique, décrivant une expérience, on souhaite extraire automatiquement de l'article les données permettant de calculer le coût équivalent carbone de l'expérience décrite. Ce coût est calculé à partir de différentes variables, tel que le temps d'apprentissage, le nombre et la nature des processeurs utilisés ...

Pour effectuer ce calcul, il faut retrouver dans le texte la valeur de ces différentes variables. Une première série d'expériences a déjà été réalisé sur ce thème, à l'aide d'un grand modèle de langage. Pour cela, on fournit au modèle de langage l'intégralité de l'article ainsi qu'une question portant sur une variable particulière.

Le problème de ce type de méthode est qu'il nécessite de grands modèles de langage, capables de prendre en compte des prompts très longs. L'idée que l'on souhaite explorer consiste à utiliser des plus petits modèles conçus pour extraire du texte les informations pertinentes.

Le projet se décompose en trois étapes :

1. génération de données synthétique. Il s'agit de générer à l'aide d'un modèle de langage du type chatGPT du texte de la longueur d'un paragraphe décrivant une expérience en spécifiant dans le prompt la valeur des différentes variables d'intérêt puis d'identifier dans le texte généré les mentions à ces variables. Ces mentions seront alors repérées à l'aide de balises.
2. entraîner à l'aide des données produites en 1, des modèles d'étiquetage (tagging), fondés sur des réseaux récurrents, du type LSTM ou GRU ou bien des modèles BERT fine-tunés sur cette tâche.
3. évaluer les modèles développés en 2 sur des données d'évaluation, plus précisément, les données développées dans le cadre du projet GREENMIR qui auront été enrichies manuellement. Plusieurs modes d'évaluation plus ou moins stricts seront proposés.

Ce sujet peut donner lieu à un stage, où l'on pourra, par exemple, générer une réponse textuelle à l'aide d'un petit modèle de langage, ou encore développer des données d'entraînement plus importantes, qui pourraient par exemple être issue d'un transfert depuis d'autres types de données.