Секция «Лингводидактика: ИКТ в обучении иностранным языкам»

Large Language Models and the Future of Science: Promises and Expectations

Научный руководитель – Щедромирская Анна

Щедромирская Анна Игоревна
Postgraduate
Московский государственный университет имени М.В.Ломоносова, Факультет иностранных языков и регионоведения, Кафедра теории преподавания иностранных языков, Москва, Россия
E-mail: annakolushkina@gmail.com

Large language models represent a recent breakthrough in artificial intelligence and the demand for LLMs is high in different spheres: from business to science.

LLM is a kind of machine learning model based on deep learning neural networks and trained on a vast corpus of text data (involving billions of words) to generate outputs for different tasks connected with natural language processing. These tasks may include text generation, sentiment analysis, content creation, question answering, reading comprehension, summarization and classification, and machine translation [1]. One of the most important characteristics is that large language models are continuously developing through training on more data and improvements in the deep learning neural networks that enable them to understand language.

Among the most promising LLMs are GPT-3 with 95 natural languages and 12 code languages, BERT with 104 languages in multilingual model, BLOOM with 46 natural languages and 13 code languages, LaMDA and many others. Such large language models are trained using a process called supervised learning. Supervised learning involves accomplishing several steps: first, a large set of text inputs and their corresponding outputs are given to the model to predict the output given a new input. The model uses an optimization algorithm to adjust its parameters to minimize the difference between its predictions and the actual outputs. Then, the training data is given to the model in small batches. The model makes predictions for each batch and changes its parameters based on the errors it sees. This process is repeated several times, allowing the model to gradually learn the relationships and patterns in the data [2].

On November 15, 2022 a new large language model called Galactica (GAL) was introduced by scientists from Meta AI. According to the researches, Galactica is a LLM aimed at automatically organizing scientific knowledge and assisting scientists [3]. It is trained on a large and curated corpus of humanity's scientific knowledge which consists of more than 48 million papers, textbooks and lecture notes, millions of compounds and proteins, scientific websites, encyclopedias and so on. The authors state that unlike existing language models, which rely on an uncurated crawl-based paradigm, Galactica's corpus is high-quality and highly curated. Galactica may be of great help to scientists because, as promised, it can summarize academic papers, solve math problems, generate Wiki articles, write scientific code, annotate molecules and proteins.

Although the expectations were quite high, the reality turned out to be prosaic. Galactica, promoted as a shortcut for researchers and students, was criticized severely by users, and the access to this LLM was closed within a couple of days after the launch. The major problem with Galactica appeared to be its incapability of telling truth from falsehood which is a basic requirement for a LLM with scientific emphasis. Users rapidly found out that Galactica generated papers and wiki articles containing misleading or completely incorrect information (for example, the history of bears in space) which is especially dangerous because it deploys the tone and structure of authoritative scientific information.

Although the recent example with GPT-3 shows that LLMs are getting much better and developing much faster these days, the area of their application remains limited due to the risks they possess.

References

1) Dilmegani C. Large Language Model Examples in 2023. 2023. URL: https://research.aimultiple.com/large-language-models-examples/
2) Dilmegani C. Large Language Model Training in 2023. 2023. URL: https://research.aimultiple.com/large-language-model-training/#top-large-language-models-by-parameter-size
3) Taylor R., Kardas M., Cucurull G. et al. Galactica: A Large Language Model for Science. 2022. URL: https://arxiv.org/abs/2211.09085