Секция 1
СОВРЕМЕННЫЕ НАПРАВЛЕНИЯ ЛИНГВИСТИЧЕСКИХ ИССЛЕДОВАНИЙ

БОЛЬШИЕ ЯЗЫКОВЫЕ МОДЕЛИ В ЛИНГВИСТИКЕ

К.В.Андренко (Брест, БрГУ имени А.С. Пушкина)

Большие данные (BD) преобразуют лингвистику, а большие языковые модели (LLM) становятся важным инструментом лингвистических исследований. LLM, обученные на огромных объемах цифровых текстов, могут обрабатывать и анализировать эти данные таким образом, который не под силу человеку традиционным методами. Они предлагают иной взгляд на язык, выявляя синтаксические и семантические связи в немаркированных наборах данных, и бросают вызов устоявшимся лингвистическим принципам. Однако их использование вызывает этические и научные проблемы, так как их понимание языка в корне отличается от человеческого. Несмотря на эти проблемы, LLM уже используются в лингвистических исследованиях и, как ожидается, помогут нам лучше понять язык.

Ключевые слова: большие данные; большая языковая модель; лингвистическое исследование; векторное представление; эмерджентность.

LARGE LANGUAGE MODELS IN LINGUISTICS

K.V.Andrenko (Brest, BrSU named after A. Pushkin)

Big Data (BD) is transforming linguistics, and Large Language Models (LLM) are becoming an essential tool for linguistic research. LLM, trained on huge volumes of digital texts, can process and analyse this data in ways that traditional methods cannot do for humans. They offer a different perspective on language, revealing syntactic and semantic relationships in unlabelled datasets, and challenge established linguistic principles. However, their use raises ethical and scientific concerns because their understanding of language is fundamentally different from human language. Despite these challenges, LLM are already being used in linguistic research and are expected to help us better understand language.

Keywords: Big Data; Large Language Model; linguistic research; vector representation; emergence.

Big Data (BD) is becoming a significant factor in the digital transformation of linguistics. The concept of BD captures data of huge volumes, diverse and connected, which are rapidly growing and changing. BD are a reflection of the explosive growth in the complexity of modern language as a cognitive and communicative toolkit for practically all types and domains of contemporary cultural and social activity. Significantly, the volume and rate of growth of big data is disproportionate to human capabilities. This is expressed, for example, in the growth of the number of texts that require analysis, the volume of special vocabulary used, etc., as well as in the growth of the number of texts to be analysed. Even the technologies of corpus linguistics, which imply direct work of researchers with search and analytical tools of language corpora, are not able to give a responsible adequate answer to the BD challenge. That is why the introduction of Large Language Models (LLM) – generative artificial neural networks trained on huge volumes of digitised textual data – into linguistic science is relevant. Big data serves as a necessary basis for

LLM, which are able to automatically process huge amounts of textual data, making them an important innovative tool for linguistic research.

LLM are based on a formal mathematical (vector) representation of texts, which differs significantly from the human representation of language. LLM are able to establish in unlabelled textual data sets many different syntactic and semantic relations and to make generalisations that may not coincide with the principles accepted in modern linguistics and expand the problem field of linguistic research. LLM actualise fundamental questions about linguistic universals and generate new problems for linguistic theories and methods. It can be assumed that LLM will help to advance scientific understanding of the very nature of language and linguistic cognition.

The first applications of LLM in linguistic research are already being formalised. For example, using LLM to test linguistic hypotheses, using linguistic knowledge to improve LLM, and developing neuro-vector-symbolic architectures to solve complex reasoning problems [1]. At the same time, LLM can themselves be considered as an object of linguistic research. For example, a core domain has been identified in LLM that corresponds to linguistic competence in 30 languages, which shows a strong correlation of linguistic competence with structure and size [2]. Studying the functional domains of LLM can help to understand the mechanisms of human linguistic abilities.

In the context of the digital transformation of the communicative environment, it should be considered that LLM may have human-level linguistic competence (but of a completely different plan) and may perform complex tasks requiring abstract knowledge and reasoning. This already raises a number of problems, which can and should be the subject of linguistic reflection. For example, N.A. Chomsky expressed his fears that the use of LLM «will lead to the degradation of science and degrade ethics by introducing a fundamentally flawed concept of language and knowledge into our technology» [3]. Chomsky argues that LLM create the illusion of meaningfulness, when in fact their operation is fundamentally different from the generation of speech by humans. According to Chomsky, LLMs cannot replace years of work by linguists to study language because they cannot replicate the instinctive process by which a child produces speech with minimal exposure to information. LLM can «extrapolate the most likely spoken response» but are unable to explain and think creatively. LLM have a different view of language, hence they must be used differently in research: for example, in research emphasising statistical latent syntactic relations. Due to their technical specificity – being tied to parameters – LLM sometimes allow us to discover linguistic patterns better and faster than humans. It is worth noting the research of Anil Ananthaswamy, where it is emphasised that LLM are not «stochastic parrots», but are capable of generalisation and creativity that goes beyond the training data (emergence) [4].

A necessarily important ontological difference between artificial intelligence and human intelligence is that a human being in the process of learning generalises facts of the real world, whereas a machine generalises texts that may have little relation to the real world.

СПИСОК ЛИТЕРАТУРЫ

1. Language models and linguistic theories beyond words [Electronic resource] // Nature. – 2023. – Mode of access: https://www.nature.com/articles/s42256-023-00703-8. – Date of access: 04.02.2024.

2. Zhao, J. Unveiling a core linguistic region in Large Language Models [Electronic resource] / J. Zhao and etc. // arXiv:2310.14928v1 [cs.CL]. – 23 Oct 2023. – Mode of access: https://arxiv.org/pdf/2310.14928.pdf. – Date of access: 01.02.2024.

3. Noam Chomsky: The False Promise of ChatGPT [Electronic resource] // International New York Times. – 9 Mar 2023. – Mode of access: https://www.nytimes.com/2023/03/08/opinion/noam-chomsky-chatgpt-ai.html. – Date of access: 21.01.2024.

4. Ananthaswamy, A. New Theory Suggests Chatbots Can Understand Text [Electronic resource] / A. Ananthaswamy // Quantamagazine [Artificial Intelligence]. – 22 Jan 2024. – Mode of access: https://www.quantamagazine.org/new-theory-suggests-chatbots-can-understand-text-20240122/ – Date of access: 29.01.2024.