УДК 81’33

Овсянникова Марина Анатольевна
кандидат филологических наук,
доцент кафедры английской филологии
Московский городской
педагогический университет
г. Москва, Россия

Marina Ovsiannikova
PhD in Philology, Associate Professor
Moscow City University
Moscow, Russia
ovsyannikovama@mgpu.ru

Николаева Марина Николаевна
кандидат филологических наук,
доцент кафедры английской филологии
Московский городской
педагогический университет
г. Москва, Россия

Marina Nikolaeva
PhD in Philology, Associate Professor
Moscow City University
Moscow, Russia
nikolaevam@mgpu.ru

НЕЙРОСЕТИ КАК ИНСТРУМЕНТ АНАЛИЗА ТЕКСТА: СРАВНИТЕЛЬНЫЙ АСПЕКТ

Данная статья представляет собой сравнительный анализ применения традиционных методов исследования языкового материала и использования «новейшего» способа обработки текста – нейронной мультимодальной сети. Традиционные методы включают в себя метод сплошной выборки, структурный анализ с целью отбора блендированных языковых единиц из массива материала, а также парсинг сайта исследуемого словаря для упрощения сбора материала. Нейросети, используемые для анализа материала, представлены на бенчмарк-платформе LMSYS Chatbot Arena. Сравнение проходит в несколько этапов: отбор материала «ручным» способом для дальнейшей верификации данных; ана-

лиз материала с помощью соответствующего промпта с дальнейшим анализом результатов на наличие галлюцинаций и ошибок. Делается вывод о несовершенстве нейросетей как инструментов анализа текста, однако подчеркивается необходимость использования новых технологий и постепенной модификации работы с ними для расширения инструментария анализа и повышения своих собственных компетенций исследователя.

К л ю ч е в ы е с л о в а: *нейронная сеть; языковая модель; генеративный предобученный трансформер (GPT); парсинг; словослияние; бленды; анализ текста.*

NEURAL NETWORK AS A TOOL OF TEXT ANALYSIS: COMPARATIVE STUDY

This article is a comparative analysis of traditional methods of studying linguistic material and the use of the "newest" method of text processing, a neural multimodal network. Traditional methods include the continuous sampling method, structural analysis which helps to select blended lexemes from the language material, as well as parsing. The neural networks used to analyze the material are presented on the LMSYS Chatbot Arena benchmark platform. The comparison takes place in several stages: the selection of the material in a “manual” way for further verification of the data; the analysis of the material using the appropriate prompt with further analysis of the results for hallucinations and errors. The article concludes that neural networks are not perfect as text analysis tools but emphasizes the need to use new technologies and gradually enhance the researcher's competencies.

K e y w o r d s: *neural network; LLM; GPT; parsing; blending; blends; text analysis.*

Междисциплинарность современной науки позволяет лингвистике наряду с традиционными методами решения прикладных задач находить применение новейшим инструментам исследования. Одним из таких инструментов, теоретически способных расширить методы анализа текста, являются нейронные сети на базе генеративного предобученного трансформера, GPT («джипити»).

Будет заблуждением сказать, что искусственный интеллект в целом и нейронные сети в частности это нечто абсолютно новое в науке, появившееся в последние несколько лет. Специалисты в области технических наук знакомы с данной сферой науки и технологии с 40–50х годов XX века, когда основные идеи начали оформляться и приобретать свою терминологию (У. МакКаллок и У. Питтс и «нейронные сети», Дж. Маккарти и «искусственный интеллект», Ф. Розенблатт и «перцептрон» и др.). Однако всеобщее погружение в эту тематику простых пользователей началось после ноября 2022 года, когда компания OpenAI объявила о создании чат-бота на основе своей нейронной сети с архитектурой генеративного предобученного трансформера ChatGPT.

Нейронные языковые модели (далее – нейросеть) обучаются на больших наборах текстовых данных, чтобы генерировать текст, схожий с человеческим. Для тренировки больших языковых моделей используются методы обучения с учителем и обучения с подкреплением. Нейронные сети – это на данный момент самый популярный инструмент машинного обучения, которое в свою очередь является самым востребованным инструментом в области искусственного интеллекта [1].

Целью настоящего исследования была попытка применить нейросети для выявления блендов, слов, образованных посредством словослияния [2], из неструктурированных данных. Неструктурированными данными называют любые данные, не имеющие заранее заданной структуры или организации. В отличие от структурированных данных, упорядоченных в удобные строки и столбцы базы данных, неструктурированные данные могут быть неотсортированной и обширной коллекцией информации [3].

Материалом исследования был выбран Dictionary of Obscure Sorrows («Словарь Странных Переживаний» – перевод наш. – *М. О., М. Н.*) [4]. Dictionary of Obscure Sorrows – это проект Джона Кенига (John Koenig) по составлению английских слов, цель которого найти неологизмы для обозначения эмоций, еще не описанных в языке. Проект словаря появился в 2006 году и первоначально был представлен в виде сайта, затем был добавлен YouTube канал, после чего последовало издание книги в 2021 году. Автор активно позиционировал свою работу в медиа, включая популярную конференцию TED, и некоторые слова из этого словаря вызвали любопытство со стороны СМИ и широкой публики. Данный материал привлек внимание авторов статьи в связи с их интересом к неологизмам в английском языке, а также к вопросу продуктивности типов словообразования. К тому же источник материала – сайт https://www.dictionaryofobscuresorrows.com/ – представляет собой несовершенный тип сайта с полимодальной информацией, неудобной навигацией и неструктурированными данными, что, по мнению авторов, является идеальным материалом для применения пока еще нового инструмента нейронной сети, которая, по словам многих специалистов, «творит волшебство».

В связи с тем, что в России доступ к некоторым нейронным моделям, в частности принадлежащим двум лидерам рынка технологий искусственного интеллекта OpenAI (ChatGPT) и Anthropic (Claude), ограничен, анализ материала был произведен на бенчмарк-платформе LMSYS Chatbot Arena [5]. Бенчмарк-платформа аккумулирует в себе все существующие и постоянно появляющиеся языковые модели, созданные на основе различных нейросетей. Существует возможность свободно пользоваться функционалом представленных нейросетей при условии, что при получении результата нейросеть объективно оценивается.

Анализ языкового материала проходил в несколько этапов. Большие языковые модели на данный момент имеют один серьезный недостаток: они склонны к конфабуляциям, или, как принято теперь называть, к галлюцинациям. При ответе на заданный вопрос нейронная сеть может дать неверный или недостаточно верный ответ. Чтобы иметь возможность верификации результатов, на первом этапе анализа был проведен парсинг сайта словаря на основе языка программирования Python с применением библиотеки BeautifulSoup. На данном этапе было выделено 103 лексические единицы в качестве леммы словарных статей, из которых 20 были выявлены как бленды.

Однако было замечено, что не все лексические единицы были отобраны с помощью инструмента парсинга. Некоторые слова вводились автором словаря как видеовставки на канале YouTube. Такая полимодальность сайта привела к необходимости подключения традиционного метода сплошной выборки, в результате чего общее количество слов было расширено до 113 единиц, 23 из которых были структурно определены как результат словослияния, бленды.

Следующий этап исследования был связан с непосредственным анализом материала с помощью запроса к нейросетям. По утверждению известного ученого в области машинного обучения Андрея Карпатого (Andrej Karpathy), с появлением больших языковых моделей естественный язык, в частности английский, становится новым языком программирования [6]. Это связано с тем, что для получения результата при работе с нейросетью необходимо правильно сформулировать промпт, то есть запрос к нейросети. Авторы статьи в полной мере осознают тот факт, что чем детальнее продуман запрос, контекст и все необходимые детали, тем лучше будет результат, однако на данном этапе нашего исследования мы обошлись простым промптом: *Find all the blended words in the following dictionary https://www.dictionaryofobscuresorrows.com/* (Найди все блендированные единицы в словаре по ссылке).

Было сделано 70 последовательных запросов на бенчмарк-платформе LMSYS Chatbot Arena в опции Arena Battle, которая представляет собой слепое сравнение двух нейросетей, название которых пользователь узнает после того, как оценил результат. Не все запросы дали результат, поскольку многие нейронные сети, представленные на платформе, либо слишком малы для выполнения такого запроса, либо предназначены исключительно для написания кода (например, reka-core-20240501).

Нейросетевые модели выдавали разный объем ответа в количестве слов, объясняя это тем, что они не предназначены для подобного анализа, или предлагали пользователю инструкцию, как самостоятельно выделить бленды в материале. В связи с чем сложно оценить количественную полноту ответа, поэтому резюме авторов по работе с нейросетями в рамках анализа неструктурированных данных с определенной лингвистической задачей носит исключительно качественный характер. Важно подчеркнуть, что в рамках нашей статьи «галлюцинациями» мы будем называть только придуманное нейросетью выражение, которое не могло быть отобрано из материала ввиду отсутствия в нем, что верифицировалось на первом этапе анализа. Лексические единицы, которые фактически содержались в материале и были приведены нейросетью как бленды, но имели другой тип словообразования, будут причислены к «ошибкам».

Оценивая эффективность выполненного запроса нейросетями платформы, на основе которой был проведен анализ языкового материала, в качестве «победителя» можно назвать gpt-4o-2024-05-13, самую новую разработку компании OpenAI. Всего было приведено 11 лексических единиц без единой галлюцинации. Однако необходимо отметить, что такие единицы, как *Jouska* и *Chrysalism*, были отнесены нейросетью к блендам, хотя они представляют примеры заимствования и аффиксации соответственно.

Еще одна нейросеть, которая обошлась без галлюцинаций, была mixtral-8x22b-instruct-v0.1, которая из 10 выданных единиц правильно определила как бленды 3 единицы.

Нейросеть компании Anthropic claude-3-sonnet-20240229 из 27 лексических единиц верно определила 8 блендов, однако 12 лексических единиц представляли собой галлюцинации. Это такие слова, как *Izzit, Exbrozzi* и *Gnosienne*. Первое слово выглядит как сокращение вопросительной формы *is it*, которое можно найти в различных текстах, а последнее напоминает название произведений для пианино, написанных французским композитором XIX века Эриком Сати “Gnossiennes”. Можно предположить, что галлюцинации являются результатом данных, на которых обучалась сеть.

Наибольшее количество слов – 35 лексических единиц – выдала нейросеть канадской компании Cohere command-r-plus. Такие единицы *kenopsia* (от греч. kenosis «emptiness» + opsia «seeing» [4]), *anecdoche* (anecdote + echo [4]), *zenosyne* (от греч. Zeno’s dichotomy paradox + Mnemosyne [4]), по мнению авторов, правомерно относятся нейросетью к блендам. Однако 20 единиц из 35 представляют собой галлюцинации с преобладанием, что интересно, слов из немецкого языка, например, *Luftgeist* и *Zeitgeber*, которые представляют собой композиты и образованы с помощью словосложения, а не словослияния. В словаре Джона Кенига действительно есть слова немецкого происхождения. Возможно, нейросеть предложила такие слова по аналогии.

Похожие галлюцинации, но из японского языка, были предложены языковой моделью llama-3-8b-instruct. Из 11 единиц 6 были настоящими словами из японского языка, однако не имели никакого отношения к анализируемому материалу (например, *Yūgen, Ikigai*). Здесь мы тоже склонны объяснять эту конфабуляцию аналогией с другими словами в словаре.

Таким образом, на данном этапе развития нейросетевых языковых моделей можно говорить о том, что использование подобного инструментария для анализа неструктурированного текстового материала должно проводиться в сочетании с традиционными «ручными» методами. В нашем случае традиционными методами стали метод сплошной выборки и метод структурного анализа в сочетании с парсингом для упрощения сбора материала. Очевидно, что использование любого инструмента анализа, не только нейросетей, возможно только при наличии у исследователя дополнительных экстраобластных, междисциплинарных, знаний для правильной расшифровки полученных результатов.

Несмотря на недостаточно удовлетворяющие запросу результаты отбора и анализа языкового материала с помощью нейросетевых инструментов, авторы статьи не намерены отказываться от использования их в работе. Работа с любой технологией требует от пользователя определенных основ технической грамотности. Однако в работе с языковыми моделями необходимо также быть компетентным в области лингвистики, чтобы уметь верифицировать и оценить полученные результаты.

ЛИТЕРАТУРА

1. Марков С. Охота на электроовец. Большая книга искусственного интеллекта. М., 2024. 568 с.

2. Афанасьева О. В., Морозова Н. Н., Антрушина Г. Б. Лексикология английского языка. English lexicology : учебник и практикум. 8-е изд., пер. и доп. М. : Юрайт, 2020. 196 с. EDN ZDRPWZ.

3. Unstructured Data: Examples, Tools, Techniques, and Best Practices [Electronic resource]. URL: https://www.altexsoft.com/blog/unstructured-data/ (accessed: 24.06.24).

4. Dictionary of Obscure Sorrows [Electronic resource]. URL: https://www.dictionaryofobscuresorrows.com/ (accessed: 24.06.24).

5. LMSYS Chatbot Arena (Multimodal): Benchmarking LLMs and VLMs in the Wild [Electronic resource]. URL: https://arena.lmsys.org/ (accessed: 24.06.24).

6. Intro to Large Language Models [Electronic resource]. URL: https://www.youtube.com/watch?v=zjkBMFhNj_g (accessed: 24.06.24).