DOI: 10.25178/nit.2024.1.1

СПЕЦИАЛЬНАЯ ТЕМА

Статья

Использование искусственного интеллекта для создания системы машинного перевода и образовательных ресурсов на тувинском языке

Марина Л. Новикова, Филипп Н. Новиков
Российский университет дружбы народов, Российская Федерация

Усовершенствование компьютерных технологий, применяемых в гуманитарных науках, прогресс в области развития больших языковых моделей, основанных на технологиях машинного обучения и нейросетей, вышел на высочайший уровень развития. Лингвистический потенциал больших языковых моделей вызывает закономерный интерес исследователей, что является обоснованным отражением актуальности и важности использования искусственного интеллекта для создания системы машинного перевода и образовательных ресурсов.

В статье рассматривается опыт создания большой языковой модели на тувинском языке с использованием машинного обучения и искусственного интеллекта. Авторами была предпринята попытка создания большой языковой модели, способной распознавать тувинский язык, осуществлять перевод фраз с тувинского языка на русский язык, или с русского на тувинский. Помимо этого, исследовались и тестировались возможности генерации текста на тувинском языке, что может быть использовано как в сфере преподавания языка, так и при проведении различных лингвистических исследований.

Актуальность исследования заключается в том, что в настоящее время тувинский язык не представлен ни в одной из известных систем машинного перевода. Важным результатом является анализ уровня цифрового присутствия тувинского языка в Интернете, а также предложенные рекомендации по выработке оптимального алгоритма построения подобных систем и веб-сервисов, основанных на машинном обучении.

Результаты исследования представляют практическую ценность не только применительно к тувинскому языку, но могут быть экстраполированы и на другие государственные языки Российской Федерации.

Ключевые слова: *тувинский язык; искусственный интеллект; машинный перевод; нейросеть; большая языковая модель; цифровое присутствие; машинное обучение*

Для цитирования:
Новикова М. Л., Новиков Ф. Н. Использование искусственного интеллекта для создания системы машинного перевода и образовательных ресурсов на тувинском языке // Новые исследования Тувы. 2024, № 1. С. 6-17. DOI: https://doi.org/10.25178/nit.2024.1.1

Новикова Марина Львовна — доктор филологических наук, профессор кафедры русского языка и лингвокультурологии Института русского языка Российского университета дружбы народов. Адрес: 117198, Российская Федерация, г. Москва, ул. Миклухо-Маклая, д. 10, корп. 3. Email: novikova-ml@rudn.ru

Новиков Филипп Николаевич — кандидат филологических наук, доцент кафедры иностранных языков юридического института Российского университета дружбы народов. Адрес: 117198, Российская Федерация, г. Москва, ул. Миклухо-Маклая, д. 6. Email: novikov_fn@pfur.ru

NOVIKOVA, Marina Lvovna, Doctor of Philology, Professor, Russian Language and Cultural Studies Department, Russian Language Institute, RUDN University. Postal address: 10, bldg. 3 Miklukho-Maklaya St., 117198, Moscow, Russia. Email: novikova-ml@rudn.ru

ORCID ID: 0000-0002-4673-067X

NOVIKOV, Philipp Nikolaevich, Candidate of Philology, Associate Professor, Law Institute Department of Foreign Languages, RUDN University. Postal address: 6 Miklukho-Maklaya St., 117198, Moscow, Russia. Email: novikov_fn@pfur.ru

ORCID ID: 0000-0003-4884-3659

SPECIAL THEME

Article

Using artificial intelligence to develop a machine translation system and teaching resources in the Tuvan language

Marina L. Novikova, Philipp N. Novikov
RUDN University, Russian Federation

The advancement of computer technologies applied in the humanities and the progress in the development of large language models based on machine learning and neural network technologies have reached an exceptionally high level of sophistication. The linguistic potential of large language models elicits a natural interest among researchers, which constitutes a justified reflection of the relevance and importance of using artificial intelligence to create machine translation systems and educational resources.

The article explores the experience of creating a large language model for the Tuvan language using machine learning and artificial intelligence. The authors undertook an attempt to develop a large language model capable of recognizing the Tuvan language, translating phrases into Russian and back. In addition, the possibilities of generating text in Tuvan were examined and tested, which can be used both in the field of language teaching and when conducting various kinds of linguistic research.

This experience is unique since, as of now, the Tuvan language is not represented in any well-established machine translation systems. A secondary aim of the research is to analyze the level of the language's digital presence on the Internet, as well as to provide recommendations for devising an optimal algorithm for building similar systems and web services based on machine learning. The research outcomes are of practical value not only with respect to the Tuvan language but can also be extrapolated to other official languages in the Russian Federation.

Keywords: *Tuvan language; artificial intelligence; machine translation; neural networks; large language models; digital presence; machine learning*

For citation:
Novikova M. L. and Novikov Ph. N. Using artificial intelligence to develop a machine translation system and teaching resources in the Tuvan language. *New Research of Tuva*, 2024, no. 1, pp. 6-17. (In Russ.). DOI: https://doi.org/10.25178/nit.2024.1.1

Введение

Тувинский язык, являющийся официальным языком Республики Тува, занимает особое место в культурном и лингвистическом пространстве Российской Федерации. Несмотря на результаты исследований, свидетельствующих о развитии асимметричного, тяготеющего к преобладанию русского языка, билингвизму в сфере школьного языкового образования (Арефьев, Бахтикиреева, Синячкин, 2021), ученые также отмечают и позитивные стороны языковой политики, подчеркивая новый дискурс языковой политики, ориентированный на поддержку тувинского языка (Боргоякова, Биткеева, 2023), повышение уровня языкового самосознания (Дырхеева, Цыбенова, 2020) и важность создания современных ресурсов с целью использования их при профессиональной подготовке работников сферы образования.

Говоря о развитии научных и образовательных проектов, посвященных тувинскому языку, очень важно упомянуть о недоступных ранее не только рядовым пользователям, но и исследовательским центрам технологиях, которые могут предоставить ученым, педагогам и лицам, интересующимся тувинским языком, абсолютно новые возможности. Одним из видов таких инновационных технологий, имеющих широкомасштабный потенциал, являются большие языковые модели (LLM — *Large Language Models*) — нейронные сети, иначе называемые искусственным интеллектом.

В 2023 г. развитие нейросетей, связанное с ростом вычислительных мощностей и повышенным вниманием, уделяемым специалистами искусственному интеллекту, привело к качественному скачку их развития. Запуск компанией Open AI сервиса ChatGPT в ноябре 2022 г., основанного на большой языковой модели GPT–3.5, способной вести диалог с пользователем практически на любую заданную тему и выполнять неограниченное количество задач, навсегда изменил смысл словосочетания «искусственный интеллект».

Под этим термином в течение нескольких прошедших десятилетий ученые понимали древовидный алгоритм, который можно было представить в виде блок-схемы. Особенно важно отметить, что искусственный интеллект был не способен выполнять те функции и приобретать те навыки, которые не были специально заложены в него человеком[1]. Эффективность использования нейросетей при создании систем машинного перевода была доказана ранее (Sreelekha et al., 2016: Электр. ресурс), все современные лидирующие сервисы машинного перевода, включая Google Translate, Яндекс Переводчик и DeepL, значительно улучшили свое качество при переходе на этот вид технологий. Кроме того, существует успешный опыт применения машинного перевода по отношению к языкам коренных народов, например, к инуитским языкам Канады (Le, Sadat, 2020).

Принимая во внимание как локальный, так и международный интерес к изучению, сохранению и популяризации тувинского языка, авторы данного исследования предприняли попытку использования новейших, ранее не использовавшихся в этом качестве, разработок в области машинного обучения и искусственного интеллекта, с целью создания системы машинного перевода и образовательных ресурсов на тувинском языке.

Цифровое присутствие тувинского языка в Интернете и потребность в обучении большой языковой модели

Согласно всестороннему исследованию Интернет-ресурсов, в которых функционирует тувинский язык, которое выполнили Ч. Г. Ондар, В. С. Донгак, Д. Ш. Монгуш (Ондар, Донгак, Монгуш, 2023), в настоящее время существует острая потребность в создании таких важных образовательных и лексикографических ресурсов, как программы и веб-сервисы для обучения тувинскому языку, специализированные словари, полноценный корпус с функционалом, позволяющим проводить статистический анализ, онлайн-переводчик, система проверки правописания, а также системы распознавания (*speech-to-text*) и голосового синтеза речи (*text-to-speech*) на тувинском языке.

Многие из этих актуальных задач могут быть частично или даже полностью решены с помощью использования нейросетей. Однако для этого требуется, чтобы нейросеть могла использовать тувинский язык для ввода и вывода данных. Эксперимент, проведенный авторами в августе 2023 г. с помощью новейшей версии языковой модели от OpenAI, а именно — GPT–4 «Generative Pre-trained Transformer» (генеративный предобученный трансформер), выявил следующий уровень поддержки некоторых языков народов России и стран СНГ (см. *таб. 1*).

Стопроцентное совпадение, указанное в приведенной *таблице 1*, означало, что языковая модель GPT-4 правильно реагировала на все команды, могла отвечать на вопросы на данном языке и безошибочно переводить абсолютно все запросы пользователя. В то время как снижение процентного соотношения свидетельствовало или об отказе системы от взаимодействия со ссылкой на недостаточные знания языка, или выдачу случайного набора слов. Важно подчеркнуть, что эти слова могли относиться как к этому же языку, так и к совершенно другому, но обычно родственному.

Сопоставление процентных соотношений позволило прийти к выводу о корреляции между уровнем цифрового присутствия языка в Интернете и качеством его поддержки со стороны модели GPT-4. Как видно из *таблицы 1*, стопроцентный показатель встречался преимущественно у языков стран СНГ. Эти языки не только предоставляют большой массив данных, доступных для обработки, но и более четко позволяют проанализировать статус каждой из веб-страниц, попавших в него, а также часто предполагают наличие достаточного количества переведенных на другие, еще более распространенные в Интернете, языки веб-страниц и текстов. Поскольку именно наличие парал-

[1] Sharma M., Muralidhar N., Ramakrishnan N. Overcoming barriers to skill injection in language modeling: Case study in arithmetic [Электронный ресурс] // arXiv preprint. 2022, 3 November. DOI: https://doi.org/10.48550/arXiv.2211.02098 (дата обращения: 05.01.2024).

лельного языкового корпуса, включающего в себя идентичные по структуре тексты, является наиболее благоприятным условием для машинного обучения, многие государственные языки республик Российской Федерации (за исключением татарского и башкирского) показали не столь высокий уровень совместимости с GPT-4. Именно этот феномен представленности тувинского языка и был описан упомянутыми выше авторами (Ондар, Донгак , Монгуш, 2023).

Таблица 1. Результаты эксперимента по уровню поддержки языков России и стран СНГ большой языковой моделью GPT-4, в %
Table 1. The results of the experiment on the level of support for the languages of Russia and the CIS countries by the large language model GPT-4, in %

Язык	*Уровень поддержки большой языковой моделью GPT-4*
Казахский	100
Азербайджанский	100
Армянский	100
Киргизский	100
Таджикский	100
Узбекский	100
Татарский	100
Башкирский	100
Туркменский	100
Белорусский	100
Якутский (саха)	45
Бурятский	40
Чеченский	35
Ингушский	35
Чувашский	35
Тувинский	**30**
Аварский	30
Кабардино-черкесский	15
Осетинский	15
Удмуртский	10

Исследование помогло выявить закономерность, согласно которой качество работы нейросети с каждым из языков прямо пропорционально количеству проиндексированных и правильно распознанных текстов на нем. Одним из ресурсов, важность роли которого в процессе машинного обучения была доказана, является Википедия (Srinivasan et al., 2021), так как ее формат обеспечивает разделение на языки и единообразную структуру статей. По данным на ноябрь 2023 г., тувинская Википедия занимает 229 место по количеству статей в энциклопедии (более 12000), находясь рядом с аварским и коми-пермяцким языками. Следует отметить, что несмотря на то, что само по себе количество статей не гарантирует высокое качество массива данных, эта статистика позволяет судить о приблизительном уровне осведомленности GPT–4 о тувинском языке, так как большие языковые модели являются не до конца описанными и малоизученными технологиями. Одна из наиболее удивительных характеристик современных больших языковых моделей, к которым относится GPT-4, заключается в том, что, обрабатывая объем, превышающий сотни миллиардов знаков, она может обучиться новым языкам, входящим в этот массив данных, вести диалог, осуществлять перевод и генерировать текст.

Проблема оценки объема массива данных, использованных для обучения большой языковой модели, заключается в том, что она представляет собой так называемый *black box*, то есть «черный ящик» — систему, в которой известны вводные и выводные данные, но нет представления о том, как именно она устроена и какие алгоритмы ею управляют[1]. Более того, массив данных, использованный для ее обучения, также остается полностью закрытым. Известно, что были использованы веб-страницы и тексты, находящиеся в открытом доступе и незащищенные авторским правом, в том числе Википедия и индексируемые поисковыми системами страницы. Однако компания OpenAI не раскрывает сведения о том, какие именно наборы информации были обработаны. Это связано, по-видимому, с возможными исками правообладателей и владельцев ресурсов, осознавших необходимость введения механизмов регулирования использования данных с целью машинного обучения.

Эти факторы делают более сложным понимание вклада каждого из ресурсов, на котором представлен тувинский язык. С развитием сначала Web 2.0, а затем и мобильных устройств, особенно способствовавших использованию специально выпущенных для них отдельных приложений, большинство пользователей создает контент на определенном языке, в том числе и на тувинском, не на индексируемых статичных веб-страницах, а на так называемом уровне *deep web* (глубокая сеть), на котором происходит обмен сообщениями в мессенджерах *WhatsApp*, *Viber* и *Telegram*, а также в большинстве популярных социальных сетей.

Чтобы продемонстрировать работу необученной модели GPT-4, сделаем следующие запросы (см. *рис. 1*).

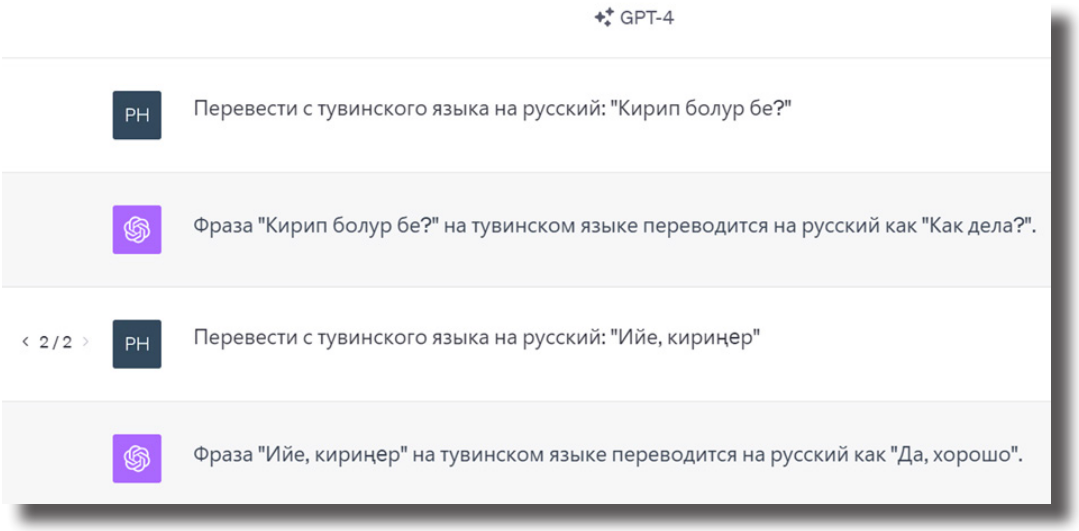

Рис. 1. Пример некорректной работы, вызванной «галлюцинациями» стандартной (необученной) модели GPT–4 (запрос ноября 2023 г.)
Fig. 1. An example of incorrect operation caused by "hallucinations" of the standard (untrained) GPT–4 model (November 2023 request)

Как видно из *рис. 1*, языковая модель не в состоянии дать правильный ответ на наш запрос. Вместо того, чтобы заявить об отсутствии данной информации в своем наборе данных (это случалось при неспособности выдать правильный ответ менее, чем в 10% случаев), она пытается «угадать» правильные ответы, которые контекстуально могли бы подойти к запросам — например, совпадает пунктуация (запятые, восклицательные знаки), но значение совершенно разное — «Кирип болур бе?» означает «Можно войти?», а «Ийе, кириңер» — «Да, войдите».

Феномен, при котором нейросеть использует случайные данные, которые либо являются неточными, либо вообще относятся к другой области знаний или другому языку, носит название «галлюцинаций» (Athaluri et al., 2023). Галлюцинации могут быть результатом случайной генерации данных. В тех ситуациях, которые интересуют нас в рамках проводимого исследования — перевод и обучение

[1] Singh C., Hsu A. R., Antonello R., Jain S., Huth A. G., Yu B., Gao J. Explaining black box text modules in natural language with language models [Электронный ресурс] // arXiv preprint. 2023, 17 May. DOI: https://doi.org/10.48550/arXiv.2305.09863 (дата обращения: 05.01.2024).

языку — некорректные ответы несут в себе особую опасность. Они часто не могут быть верифицированы пользователем, изучающим язык или пытающимся осуществить перевод либо с тувинского языка, либо на тувинский.

Проведя анализ влияния цифрового присутствия тувинского языка на качество массива данных и выявив критически важные ограничения в существующей системе искусственного интеллекта, можно прийти к выводу о том, что требуется создание отдельной специализированной языковой модели, изначально направленной на использование тувинского языка. Это послужило основой необходимости проведения нашего эксперимента в рамках данного исследования.

Эксперимент по использованию искусственного интеллекта для создания системы машинного перевода с тувинского языка на русский

Особенно следует отметить, что поскольку даже новейшая и доступная ограниченному количеству пользователей языковая модель GPT–4 не поддерживает тувинский язык в достаточном для обучающих и переводческих целей объеме, для работы с ним требуется создание уникальной большой языковой модели, специально обученной тувинскому языку. Как уже было сказано выше, именно рост вычислительных мощностей послужил основным фактором, способствующим развитию искусственного интеллекта. Обучение качественной большой языковой модели с нуля не только требует большого, тщательно отобранного и правильно отформатированного массива данных, но и наличия высокотехнологичного оборудования, недоступного рядовым пользователям. В то время как создание собственной большой языковой модели, поддерживающей тувинский язык, представляется возможным при надлежащем финансировании или при появлении нового оборудования.

В рамках проведенного нами исследования был использован новейший сервис, представленная компанией OpenAI в сентябре 2023 г. *fine-tuning* — модификация существующей языковой модели GPT-3.5 путем её дополнительного обучения. Этот процесс осуществляется через доступ к API (*application programming interface* — интерфейс программирования). Преимущество такого способа машинного обучения состоит в том, что он не требует больших вычислительных мощностей и устанавливается не на локальный сервер пользователя, а на сервер OpenAI.

На *рисунке 2* приведена схема использованного нами процесса дополнительного обучения большой языковой модели.

Рис. 2. Схема дополнительного обучения большой языковой модели GPT–3.5 с помощью процесса fine-tuning (модификации)
Fig. 2. The scheme of additional training of the GPT–3.5 large language model using the fine-tuning process (modifications)

Для адаптивного процесса *fine-tuning* использовалось облачное пространство Google Colab, в котором запускалась программа на языке Python, позволявшая серверу Google связаться с сервером OpenAI, на котором установлена модель GPT–3.5, и добавить туда массив данных на тувинском и русском языках.

В качестве массива данных использовался неспециализированный параллельный корпус предложений бытовой тематики, в котором предварительно нужно было произвести разметку в формате .jsonl — JSON Lines (JavaScript Object Notation Lines) по следующему образцу:

{«messages»: [{«role»: «system», «content»: «Ты переводчик с тувинского языка на русский.»}, {«role»: «user», «content»: «Четтирдим, эки-дир»}, {«role»: «assistant», «content»: «Спасибо, хорошо»}]}

В приведенном выше примере после {«role»: «system», «content»} указывается роль, которая отводится искусственному интеллекту в рамках данного проекта. Она может быть общей или более узконаправленной, например, быть посвященной созданию учебных материалов, поддержанию диалога на тувинском языке, лингвистическому анализу и т. д.

Вторая часть строки кода {«role»: «user», «content»} содержит предполагаемый запрос со стороны пользователя — в данном случае, фразу на тувинском языке. Последняя часть {«role»: «assistant», «content»} заключает в себе ответ на русском языке.

Важно отметить, что для обучения системы грамматике, идиоматическим выражениям и другим аспектам, присущим живому языку, необходимо предоставлять языковой модели именно синтагмы, а не словарные формы. Так как модель GPT–3.5 уже имеет правильное представление о синтаксических и морфологических категориях русского языка, ей необходимо предоставить как можно более диверсифицированный набор данных, включающий максимальное количество минимальных пар.

После формирования массива данных, его правильной разметки и загрузки в Google Colab, был запущен процесс *fine-tuning*. Обучение модели GPT–3.5 происходит по алгоритму, называемому *few-shot learning* (Garcia et al, 2023) — обучение моделей с ограниченным количеством размеченных данных. Так как механизм *fine-tuning* до сих пор является новым и используется в экспериментальном режиме, даже после формирования массива данных потребовалось несколько попыток обучения, чтобы получить функциональную модель. Одним из решающих факторов, отделивших неэффективную модель от эффективной, стало количество «эпох», т. е. этапов обучения.

На *рисунке 3* приведены результаты третьей и четвертой попыток обучения моделей GPT–3.5 тувинскому языку — первая из них прошла в 3 эпохи.

Рис. 3. Результат обучения большой языковой модели GPT–3.5 Turbo (3 эпохи обучения)
Fig. 3. The result of learning a large GPT–3.5 Turbo language model (3 epochs of learning)

В результате получившаяся после трех эпох обучения модель несмотря на то, что это число было выбрано программой автоматически, выдавала описанные выше галлюцинации в виде случайных результатов. Более того, они уступали по релевантности и точности тем результатам, которые предоставляла модель GPT–4 по умолчанию, так как последняя заранее была обучена на гораздо большем количестве параметров.

Удачной попыткой обучения стала четвертая, при которой был надлежащим образом отформатирован массив данных, была правильным образом задана роль искусственного интеллекта, а количество эпох составило 6, а не 3 (*рис. 4*).

Рис. 4. Результат обучения большой языковой модели GPT–3.5 Turbo (6 эпох обучения)
Fig. 4. The result of learning a large GPT–3.5 Turbo language model (6 epochs of learning)

Как видно из *рис. 4*, параметр Training Loss (потеря данных во время обучения) при обучении за 6 эпох составил *0.0764,* что на порядок меньше, чем показатель *0.8076* при обучении за 3 эпохи. Следует отметить, что отмеченный выше закрытый характер системы не позволяет заранее определить оптимальное количество эпох обучения, так как может привести к нежелательному феномену, называемому *overtraining* (излишнее обучение). В таком случае языковая модель становится слишком узконаправленной и впоследствии не может выйти за рамки тех данных, которые были заложены в нее заранее (Armstrong et al., 2022).

Результатом наиболее успешной попытки обучения, выполненной нами в рамках данного исследования, явился тувинско-русский переводчик, работающий через диалоговое окно API (*рис. 5*).

Ограниченный объем данных, на котором была обучена система, позволял комбинировать лексику и корректно переводить новые высказывания, не загруженные в систему заранее. Полученные результаты, представленные на *рисунке 6* позволяют сделать вывод о корректной реализации поддержки современного алфавита тувинского языка и отсутствии ограничений на употребление и распознавание букв “ң”, “ө” и “ү”, вместо которых при проблемах с кодировкой на экран выводились

Рис. 5. Пример работы системы машинного перевода, использующей обученную в рамках эксперимента fine-tuned модель GPT–3.5 Turbo.
Fig. 5. An example of the operation of a machine translation system using the GPT-3.5 Turbo model trained in the framework of the fine–tuned experiment.

бы другие символы (Папын, 2010). Их наличие в большой языковой модели и верный ввод и вывод являются результатом обучения на чрезвычайно большом по объему массиве данных, в который также попала информациях на этих языках. Эти данные свидетельствуют об отлаженности механизма обучения искусственного интеллекта тувинскому языку с абсолютно новым для системы алфавитом.

Следует также упомянуть о том, что параметр *temperature* (температура) в настройках модели позволяет увеличить вариативность текста. При параметре, равном единице, выдавался наиболее близкий к изначально заданному результат, но разнообразие может быть полезным при применении данной модели в других целях.

Таким образом, эксперимент, успешно проведенный в рамках нашего исследования, показал, что обучение большой языковой модели тувинскому языку с помощью *fine-tuning* является возможным и, скорее всего, станет еще более доступным в ближайшем будущем.

Возможности дальнейшего использования искусственного интеллекта для создания ресурсов и сервисов на тувинском языке

Используемый нами механизм обучения языковой модели тувинскому языку предполагает наличие у разработчиков доступа к API. Это означает, что на его основе уже сейчас можно создать отдельный сайт, находящийся в доменной зоне ru, доступный всем желающим, обеспечивающий машинный перевод с тувинского языка на русский и с русского на тувинский язык с применением нейросети.

Кроме проанализированных возможностей, нейросеть можно обучить с целью генерации учебных текстов на тувинском языке и задать иную роль искусственного интеллекта, используя способ форматирования массива данных, описанный выше. В ближайшем будущем эти технологии также могут упростить создание *Telegram*-ботов на тувинском языке, языковых сервисов и обучающих приложений с элементами геймификации для различных сегментов аудитории.

Подчеркнем еще раз, что обзор информационных, коммуникационных и справочных интернет-ресурсов, содержащих контент на тувинском языке с учетом ситуации, описанной Ч. Г. Ондар, В. С. Донгак, Д. Ш Монгуш (Ондар, Донгак, Монгуш, 2023), свидетельствует о насущной потребности в большом количестве контента на тувинском языке. Этим процессам может и должен способствовать искусственный интеллект. В особенности это касается современных средств массой информации, популярных среди молодого поколения. Искусственный интеллект может осуществить локализации интерфейса приложений и различных порталов. Это может быть выполнено с помощью модели *crowdsourced translation* (перевод силами сообщества), с опорой на помощь волонтеров, желающих совершенствовать перевод на добровольной основе (Zwischenberger, 2022).

Количество возможностей в сфере образования крайне велико, так как нейросеть может облегчить не только создание традиционных обучающих материалов, включающих в себя учебники, рабочие тетради, хрестоматии и глоссарии, но и в сочетании с другими технологиями может послужить основой для создания аудиовизуальных материалов, интерактивных курсов и даже виртуального учителя, реагирующего на запросы пользователя и объясняющего особенности тувинского языка и культурные реалии Тувы.

Контроль над обучением и дальнейшим функционированием модели со стороны человека крайне важен и ни в коем случае не отменяется даже при высокой степени ее качества, так как необходимо заранее определить цели обучения модели. В настоящее время существующие параллельные базы данных являются достаточно разрозненными и неоднородными, поэтому следует внимательно отнестись к фильтрации данных уже на самом первом этапе.

При построении большой языковой модели и дальнейшем ее использовании следует также учитывать проблемы перевода концептов культуры с одного языка на другой. Исследования, проведенные в рамках языковой пары «тувинский язык — русский язык», показали, что даже при наивысшем уровне владения обоими языками и глубоких познаниях об обеих культурах, даже если переводы выполняют билингвами-редакторами, возникают определенные проблемы, которые могут ввести читателя или собеседника в заблуждение, или же передать информацию лишь частично (Кужугет, Сувандии, Ламажаа, 2021).

Помимо непосредственно языкового фактора ограничений, связанных с передачей информации, несущей в себе этнокультурный компонент, речь идет также и об ограничениях непосредственно в

сфере машинного обучения (Spennemann, 2023). Эти обстоятельства обусловливают особую важность лингвистических задач, которые могут быть решены путем создания определенного набора языковых и культурных реалий, а также последующего тестирования их восприятия и применения искусственным интеллектом. Тот факт, что нейросеть GPT была обучена в первую очередь на индоевропейских языках, также накладывает дополнительный отпечаток на ее социокультурную картину мира. В определенной мере эти проблемы могут быть решены, опираясь на принципы *Indigenous methodology*, которая предусматривает этичное, уважительное и эмпатическое исследование культуры (Тувинцы. Родные … , 2022).

Заключение

Опыт модификации предобученной большой языковой модели GPT–3.5 Turbo с помощью механизма *fine-tuning*, проведенный нами, показал возможность подобного подхода и позволил описать перспективность его применения к тувинскому языку.

Наиболее совершенная, но, как показал результат нашего эксперимента, все равно нуждающаяся в серьезной доработке, большая языковая модель GPT–4 недоступна для дополнительного обучения с помощью механизма *fine-tuning*, но ожидается, что данная функция будет доступна в ближайшем будущем.

Кроме того, GPT–4, обучение которой, по данным её разработчиков, ограничено 2021 г.[1], продолжает обучаться, используя данные, предоставляемые различными пользователями во время диалогов с ChatGPT. Таким образом, существует теоретическая вероятность того, что при условии интенсивного использования тувинского языка как сервис OpenAI, так и другие языковые модели смогут повысить уровень, на котором поддерживается тувинский язык. Однако, это обучение является неконтролируемым, и невозможно ожидать, что популярная языковая модель, в которую заранее не была заложена критическая масса данных на тувинском языке, будет его корректно использовать.

Еще одним важным аргументом в пользу создания отдельных языковых моделей на тувинском языке является следующий: на дальнейшее развитие искусственного интеллекта может повлиять критическое отношение к развитию больших языковых моделей, ведущее не только к регулированию и удалению данных, нарушающих авторские права и этические нормы, но и к сопутствующей фильтрации уже занесенной в массив данных полезной информации, в том числе информации на языках, официально не поддерживающихся языковой моделью.

В заключение следует подчеркнуть, что возможности использования искусственного интеллекта применительно к тувинскому языку безграничны, но требуют особого внимания и бережного отношения к культурному наследию и его будущему. Также хотелось бы выразить надежду на то, что проекты, основанные на подобных языковых моделях, получат дальнейшее распространение и будут развиваться как по инициативе различных организаций, так и силами энтузиастов — ценителей тувинского языка и культуры Тувы во всем мире.

СПИСОК ЛИТЕРАТУРЫ

Арефьев, А. Л., Бахтикиреева, У. М., Синячкин, В. П. (2021) Проблемы билингвизма в системе школьного языкового образования Республики Тыва // Новые исследования Тувы. №. 1. С. 255–272. DOI: https://doi.org/10.25178/nit.2021.1.14

Боргоякова, Т. Г., Биткеева, А. Н. (2023) Тувинский компонент билингвального пространства или размышления о стратегии государственной поддержки тувинского языка // Новые исследования Тувы. № 4. С. 290–300. DOI: https://doi.org/10.25178/nit.2023.4.20

Дырхеева, Г. А., Цыбенова, Ч. С. (2020) Языковые установки и языковая лояльность носителей малых языков в условиях национально-русского двуязычия (на примере бурят и тувинцев) // Новые исследования Тувы. № 1. С. 62–74. DOI: https://doi.org/10.25178/nit.2020.1.5

Кужугет, Ш. Ю, Сувандии, Н. Д., Ламажаа, Ч. К. (2021) Проблемы перевода концептов культуры на другой язык (на примере тувинских концептов культуры // Полилингвиальность и транскультурные практики. Т. 18. № 4. С. 405–420. DOI: https://doi.org/10.22363/2618-897X-2021-18-4-405-420

[1] GPT-4 [Электронный ресурс] // OpenAI. URL: https://openai.com/research/gpt-4 (дата обращения: 05.01.2024).

Ондар, Ч. Г., Донгак, В. С., Монгуш, Д. Ш. (2023) Тувинский язык в Интернете: представленность, проблемы и перспективы // Новые исследования Тувы. № 1. С. 186–207. DOI: https://doi.org/10.25178/nit.2023.1.11

Папын, А. С (2010) Тувинская раскладка клавиатуры // Новые исследования Тувы. № 1. С. 19–25.

Тувинцы: родные люди (2022) / Ламажаа Ч. К., Сувандии Н. Д., Кужугет Ш. Ю., Майны Ш. Б., под ред. Ч. К. Ламажаа, Н. Д. Сувандии. СПб. : Нестор-История. 344 с.

Athaluri, S. A., Manthena, S. V., Kesapragada, V. K. M., Yarlagadda, V., Dave, T., Duddumpudi, R. T. S. (2023) Exploring the boundaries of reality: investigating the phenomenon of artificial intelligence hallucination in scientific writing through ChatGPT references // Cureus. № 15 (4). DOI: https://doi.org/10.7759/cureus.37432

Armstrong, L. E., Bergeron, M. F., Lee, E. C., Mershon, J. E., & Armstrong, E. M. (2022) Overtraining syndrome as a complex systems phenomenon // Frontiers in Network Physiology. № 1 (20). DOI: https://doi.org/10.3389/fnetp.2021.794392

Garcia, X. Bansal, Y, Cherry, C., Foster, G., Krikun, M., Feng, F., Johnson, M., First, O. (2023) The unreasonable effectiveness of few-shot learning for machine translation // International Conference on Machine Learning. PMLR. P. 10867–10878. DOI: https://doi.org/10.48550/arXiv.2302.01398

Le, T. N., Sadat, F. (2020) Revitalization of indigenous languages through pre-processing and neural machine translation: The case of Inuktitut // Proceedings of the 28th International Conference on Computational Linguistics. P. 4661–4666. DOI: https://doi.org/10.18653/v1/2020.coling-main.410.

Sreelekha, S., Bhattacharyya, P., Jha, S. K., Malathi, D. (2016) A survey report on evolution of machine translation [Электронный ресурс] // IJCTA, 9 (33), pp. 233–240. URL: https://www.serialsjournals.com/abstract/65435_article-24.pdf (дата обращения: 12.11.2023).

Srinivasan, K., Raman, K., Chen, J., Bendersky, M., Najork, M. (2021) Wit: Wikipedia-based image text dataset for multimodal multilingual machine learning // Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval. P. 2443–2449. DOI: https://doi.org/10.48550/arXiv.2103.01913

Spennemann, D. H. R. (2023) ChatGPT and the generation of digitally born “knowledge”: How does a generative AI language model interpret cultural heritage values? // Knowledge. № 3. P. 480–512. DOI: https://doi.org/10.3390/knowledge3030032

Zwischenberger, C. (2022) Online collaborative translation: its ethical, social, and conceptual conditions and consequences // Perspectives. № 30 (1). P. 1–18. DOI: https://doi.org/10.1080/0907676X.2021.1872662

Дата поступления: 12.01.2023 г.

Дата принятия: 12.02.2024 г.

REFERENCES

Arefyev, A. L., Bakhtikireeva, U. M. and Sinyachkin, V. P. (2021). Issues of bilingualism in the school language education system of the Republic of Tuva. *New Research of Tuva*, no. 1, pp. 255–272. (In Russ.) DOI: https://doi.org/10.25178/nit.2021.1.14

Borgoiakova, T. G. and Bitkeeva, A. N. (2023) The Tuvan component of the bilingual space or reflections on the strategy of state support of the Tuvan language. *New Research of Tuva*, no. 4, pp. 290–300. (In Russ.) DOI: https://doi.org/10.25178/nit.2023.4.20

Dyrkheeva, G. A. and Tsybenova, Ch. S. (2020) Language attitudes and language loyalty of minor language speakers under the conditions of national-Russian bilingualism: the case of Buryats and Tuvans. *New Research of Tuva*, no. 1, pp. 62–74. (In Russ.) DOI: https://doi.org/10.25178/nit.2020.1.5

Kuzhugget, Sh. Yu., Suvandii, N. D. and Lamazhaa, Ch. K. (2021) The problem of translating cultural concepts into another language (on the example of Tuvan cultural concepts). *Polylinguality & transcultural practices*, no. 18. (4), pp. 405–420. (In Russ.) DOI: https://doi.org/10.22363/2618-897X-2021-18-4-405-420

Ondar, Ch. G., Dongak, V. S. and Mongush, D. Sh. (2023). The Tuvan language on the Internet: representation, challenges, and prospects. *New Research of Tuva*, no. 1, pp. 186–207. (In Russ.) DOI: https://doi.org/10.25178/nit.2023.1.11

Papyn, A. S. (2010) Tuvan keyboard layout. *New Research of Tuva*, no. 1, pp. 19–25. (In Russ.)

Tuvans: Native People (2022). Ed. by Ch. K. Lamazhaa and N. D. Suvandii. St. Petersburg, Nestor-Istoriya. 344 pp. (In Russ.).

Athaluri, S. A., Manthena, S. V., Kesapragada, V. K. M., Yarlagadda, V., Dave, T. and Duddumpudi, R. T. S. (2023) Exploring the boundaries of reality: investigating the phenomenon of artificial intelligence hallucination in scientific writing through ChatGPT references. *Cureus*, no. 15(4). DOI: https://doi.org/10.7759/cureus.37432

Armstrong, L. E., Bergeron, M. F., Lee, E. C., Mershon, J. E. and Armstrong, E. M. (2022) Overtraining syndrome as a complex systems phenomenon. *Frontiers in Network Physiology*, no. 1 (20). DOI: https://doi.org/10.3389/fnetp.2021.794392

Garcia, X. Bansal, Y, Cherry, C., Foster, G., Krikun, M., Feng, F., Johnson, M. and First, O. (2023) The unreasonable effectiveness of few-shot learning for machine translation. *International Conference on Machine Learning*, PMLR, pp. 10867-10878. DOI: https://doi.org/10.48550/arXiv.2302.01398

Le, T. N. and Sadat, F. (2020) Revitalization of indigenous languages through pre-processing and neural machine translation: The case of Inuktitut. *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 4661-4666. DOI: https://doi.org/10.18653/v1/2020.coling-main.410.

Sreelekha, S., Bhattacharyya, P., Jha, S. K. and Malathi, D. (2016) A survey report on evolution of machine translation. *IJCTA*, 9 (33), pp. 233–240 [online]: https://www.serialsjournals.com/abstract/65435_article-24.pdf (access date: 12.11.2023).

Srinivasan, K., Raman, K., Chen, J., Bendersky, M. and Najork, M. (2021) Wit: Wikipedia-based image text dataset for multimodal multilingual machine learning. *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 2443–2449. DOI: https://doi.org/10.48550/arXiv.2103.01913

Spennemann, D. H. R. (2023) ChatGPT and the generation of digitally born “knowledge”: How does a generative AI language model interpret cultural heritage values? *Knowledge*, no. 3, pp. 480–512. DOI: https://doi.org/10.3390/knowledge3030032

Zwischenberger, C. (2022) Online collaborative translation: its ethical, social, and conceptual conditions and consequences. *Perspectives*, no. 30 (1), pp. 1–18. DOI: https://doi.org/10.1080/0907676X.2021.1872662

Submission date: 12.01.2023.

Accepted date: 12.02.2024.