УДК [009+81]:004.9
DOI:10.17853/2587-6910-2022-05-53-57

ЦИФРОВАЯ ИСТОРИЯ И ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ: ПЕРСПЕКТИВЫ И РИСКИ ПРИМЕНЕНИЯ БОЛЬШИХ ЯЗЫКОВЫХ МОДЕЛЕЙ

DIGITAL HISTORY AND ARTIFICIAL INTELLIGENCE: PERSPECTIVES AND RISKS OF PRETRAINED LANGUAGE MODELS

Алексей Валерьевич Кузнецов
научный сотрудник
historyras@gmail.com
Институт всеобщей истории РАН, Москва, Россия

Alexey Valer'evich Kuznetsov
Institute of World History of Russian Academy of Sciences, Moscow, Russia

Аннотация. *Анализируется опыт создания языковых моделей на основе трансформеров для исторических языков, поскольку текстовые данные являются базой для большинства исторических исследований, что делает особенно значимым для развитие методов и технологий обработки естественного языка исторической науки. Рассмотрены возможные риски и перспективы внедрения подобных языковых моделей.*

Ключевые слова: *цифровая история, цифровая гуманитаристика, обработка естественного языка, машинное обучение/*

Abstract. *Textual data is the basis for most of historical researches. This circumstance makes the development of methods and technologies of natural language processing especially significant for historical science. In recent years, deep learning methods have dominated the field of natural language processing. Many variants of large pre-trained language models have emerged. This article analyzes the experience of creating language models based on transformers for historical languages. Possible risks and prospects for their implementation are considered.*

Keywords: *digital history, digital humanities, natural language processing, machine learning.*

Двадцать первое столетие для социальных и гуманитарных наук стало веком цифрового поворота, который вызвал к жизни новое понимание цифровой гуманитаристики, как перспективного междисциплинарного направления, объединяющего методику традиционных гуманитарных наук с компьютерными науками. Не осталась в стороне от этих процес-

сов и историческая наука, в рамках которой развивается такое направление, как цифровая история. В настоящее время еще не выработано однозначного определения ни цифровой гуманитаристики, ни цифровой истории [1]. Согласно наиболее общему представлению, под цифровой историей понимается научное направление в рамках исторической науки, связанное с использованием цифровых медиа, информационных технологий, количественных методов и методов науки о данных в практике исторических исследований, в задачах организации, презентации и визуализации исторической информации, в историческом образовании [2]. Как и историческая наука, цифровая история имеет дело с разнообразными источниками, но основными по-прежнему остаются текстовые. Данное обстоятельство делает особенно существенным для цифровой истории развитие методов и технологий обработки естественного языка (англ. *Natural Language Processing*) — направлений искусственного интеллекта и математической лингвистики, изучающих проблемы компьютерного анализа и синтеза текстов на естественных языках.

Обработка естественного языка является важнейшей составляющей широкого спектра программных приложений, которые мы используем в повседневной жизни. Приложениями на ее основе буквально наполнены современные смартфоны. К наиболее типичным задачам обработки естественного языка относятся автоматический перевод с одного языка на другой, поиск информации в текстовых корпусах, распознавание речи, анализ текста и автоматическая генерация текста. В последние годы в области обработки естественного языка стали преобладать методы глубокого обучения, показавшие чрезвычайную эффективность в решении этих задач. А среди глубоких нейронных сетей абсолютными рекордсменами в скорости и качестве работы с текстами в настоящее время являются трансформеры — нейросети, ориентированные на обработку последовательностей с использованием механизма внимания.

В данной статье мы хотим проанализировать существующий опыт создания языковых моделей на основе трансформеров для исторических языков, охарактеризовать современный взгляд на перспективы и риски широкого внедрения таких моделей.

Впервые трансформеры были описаны в 2017 г. инженерами Google Brain в работе «Attention Is All You Need» [3]. С момента появления трансформеры обрели небывалую популярность и ныне используются во множестве сервисов. Мы не будем останавливаться на технических аспектах работы трансформеров, их архитектура неоднократно описана в специализированной литературе [4]. Обратим внимание на два важных для нас факта. Во-первых, трансформеры позволили создать языковые модели на основе обработки действительно огромного объема текстовых данных. Эти модели значительно увеличили качество выполнения разнообразных задач обработки естественного языка, но процесс создания таких моделей необычайно трудоемкий и дорогостоящий. Во-вторых, уже обученные модели могут с минимальными затратами времени и труда быть оптимизированы под конкретные задачи с использованием метода переноса знаний (англ. *transfer learning*). В настоящее время не сложилось еще однозначного наименования подобных языковых моделей. Их называют большими предварительно обученными моделями, фундаментальными моделями, контекстуальными моделями и нейронными моделями.

Одной из самых первых моделей на основе трансформеров стала языковая модель BERT (англ. *Bidirectional Encoder Representations from Transformers*) компании Google, разработанная в 2018 г. Англоязычная модель была обучена на статьях из Википедии, содержащих более 2,5 млрд слов. В 2019 г. была выпущена многоязычная версия BERT, в последней версии поддерживающая 104 языка.

Другой популярной языковой моделью является GPT (англ. *Generative Pre-trained Transformer*) компании OpenAI. Последняя версия — GPT-3 — обучена на наборе данных объемом 570 Гб, а общее количество ее параметров составляет 175 млрд. GPT-3 может вести с пользователем вполне осмысленный диалог, генерировать тексты, производить семантический поиск. На базе GPT-3 разработан помощник для автоматического написания кода GitHub Copilot. По некоторым оценкам с момента релиза Copilot в августе 2021 г. 30 % нового кода в репозиториях GitHub написано с его помощью.

В настоящее время весь потенциал трансформеров еще не раскрыт в полной мере. Помимо обработки естественного языка они все больше используются и в других задачах, например, в таких, как компьютерное зрение. Различные варианты языковой модели BERT от Google служат основой для разработки новых моделей, в том числе и адаптированных для исторических языков. Сейчас уже обучены языковые модели для многих современных языков, среди которых есть и русский [5]. Оптимизированные под конкретные языки модели демонстрируют существенное улучшение качества выполнения задач обработки естественного языка по сравнению с базовой моделью BERT [6].

Успехи больших предварительно обученных моделей не могли не обратить на них внимание гуманитариев, в том числе историков, стремящихся использовать потенциал этих моделей для работы с текстами исторических источников. Так, международный открытый проект Universal Dependencies на основе мультиязычного вариант BERT и собственных трибанков (семантически и морфологически аннотированных текстовых корпусов) создал предварительно обученные модели для множества языков. В настоящее время в версии 2.6 доступны 99 моделей для 63 языков [7]. Помимо моделей для современных языков, есть модели для древнегреческого, латинского, готского, древнерусского, церковнославянского языков. Пока модели доступны только в виде онлайн-сервиса. По заверению авторов, в скором времени их можно будет скачивать. Все модели демонстрируют улучшение качества работы, по сравнению с моделями предыдущего поколения.

Среди специализированных языковых моделей назовем Latin BERT — модель для латинского языка, обученную на основе корпуса латинских текстов, включающих в себя 642,7 млн слов из различных источников от классической эпохи до неолатинских текстов XXI в. [8]. Для английского языка раннего нового времени создана модель MacBERTh, обученная на большом корпусе англоязычных текстов, написанных за время с 1450 по 1950 гг., общим объемом 3,9 млрд слов [9]. Имеется модель и для английского языка XIX в. [10].

Широкое распространение больших предварительно обученных моделей помимо очевидных преимуществ, породило новые проблемы и риски. Так, создатели GPT-3 отказываются выкладывать свою модель в открытый доступ, поскольку опасаются, что она может быть использована для массовых спам-атак, генерации экстремистских текстов, кампаний по дезинформации. За два последних года в академической среде был поднят во-

прос о рисках внедрения больших языковых моделей. В сентябре 2020 г. в публикации исследователей из колледжа в Миддлбери [11] было высказано опасение о возможной радикализации общества в случае использования преимуществ больших языковых моделей, демонстрирующих значительные успехи в генерации экстремистских текстов по сравнению с моделями предыдущих поколений.

В марте 2021 г. в статье «*On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?*» [12] был сформулирован целый комплекс проблем создания и применения больших языковых моделей. По мнению авторов, обучение таких моделей связано с неоправданными экологическими (большой объем энергии, необходимый для обучения модели) и финансовыми издержками. Стремление создателей заложить в модель все больше языковых данных приводит к тому, что в их числе оказываются национальные и гендерные стереотипы, конспирологические теории, экстремистские материалы. Поэтому при подготовке моделей особенно подчеркивается необходимость тщательно курировать наборы обучающих данных. Для моделей, обученных на выборке текстов из интернета, была и остается актуальной проблема наличия в них конфиденциальной информации [13].

В августе 2021 г. большой коллектив исследователей подготовил 200-страничный отчет о перспективах и рисках внедрения больших предобученных моделей [14]. В выводах авторы подчеркивают, что в настоящее время большие модели только начали трансформировать способы создания и развертывания систем искусственного интеллекта в мире. Их потенциал не только до конца не раскрыт, но даже не очерчен. Дальнейшая ответственная разработка и внедрение этих моделей должны быть основаны на сотрудничестве между учеными из различных областей знания, способными дать оценку моделям не только с технической, но и с этической и гуманитарной сторон.

Согласно данным социологических опросов Россия является страной технооптимистов. Большинство россиян (63 % из опрошенных в 2020 г.) считают, что наука оказывает положительное влияние на жизнь людей [15, с. 268–271]. Люди верят в технический прогресс. Поэтому, вероятно, некоторые риски, возникающие с развитием больших предобученных языковых моделей, им могут показаться надуманными.

В заключение отметим следующее: если исключить возможное использование таких моделей радикалами и криминалитетом, то для академических ученых, в том числе и историков, более значимы, на наш взгляд, их преимущества. И в первую очередь это касается качества выполнения задач обработки естественного языка. Что открывает новые перспективы и возможности при анализе текстов, в том числе и исторических. Не вызывает сомнения также и то, что уже в ближайшее время появятся новые модели, адаптированные для языков исторических источников прошлых веков, а историки с интересом будут осваивать новые инструменты и методы анализа этих источников.

Список литературы

1. *Гарскова, И. М.* «Цифровой поворот» в исторических исследованиях: долговременные тренды / И. М. Гарскова. Текст: электронный // Историческая информатика. 2019. № 3 (29). С. 57–75. URL: https://nbpublish.com/library_read_article.php?id=31251.

2. *Бородкин, Л. И.* Digital History: применение цифровых медиа в сохранении историко-культурного наследия? / Л. И. Бородкин. Текст: электронный // Историческая информатика. 2012. № 1 (1). С. 14–21. URL: https://istina.msu.ru/publications/article/2697103/.

3. *Attention* is All you Need. URL: https://www.cs.ubc.ca/~lsigal/532S_2018W2/3c.pdf. Text: electronic.

4. *Rothman, D.* Transformers for Natural Language Processing: Build innovative deep neural network architectures for NLP with Python, PyTorch, TensorFlow, BERT, RoBERTa, and more / D. Rothman. Birmingham: Packt Publishing Ltd., 2021. 360 p. URL: https://www.rulit.me/data/programs/resources/pdf/Transformers-for-Natural-Language-Processing_RuLit_Me_661922.pdf. Text: electronic.

5. *Kuratov, Yu.* Adaptation of deep bidirectional multilingual transformers for Russian language / Yu. Kuratov, M. Arkhipov. Текст: электронный // Компьютерная лингвистика и интеллектуальные технологии. 2019. Вып. 18 (25). С. 333–339. URL: https://publications.hse.ru/mirror/pubs/share/direct/284412707.

6. *Nozza, D.* What the [mask]? making sense of language-specific BERT models / D. Nozza, F. Bianchi, D. Hovy. https://doi.org/10.48550/arXiv.2003.02912. Text: electronic.

7. *UDPipe* 2 Models. Text: electronic // Institute of Formal and Applied Linguistics. URL: https://ufal.mff.cuni.cz/udpipe/2/models.

8. *Bamman, D.* Latin bert: A contextual language model for classical philology / D. Bamman, P. J. Burns. https://doi.org/10.48550/arXiv.2009.10053. Text: electronic.

9. *Manjavacas, E.* MacBERTh: Development and Evaluation of a Historically Pre-trained Language Model for English (1450–1950) / E. Manjavacas, L. Fonteyn. Text: print // Proceedings of the Workshop on Natural Language Processing for Digital Humanities (NLP4DH). Stroudsburg, 2021. P. 23–36.

10. *Neural* Language Models for Nineteenth-Century English / K. Hosseini, K. Beelen, G. Colavizza, M. C. Ardanuy. Text: electronic // Journal of Open Humanities Data. 2021. Vol. 7. http://doi.org/10.5334/johd.48.

11. *McGuffie, K.* The radicalization risks of GPT-3 and advanced neural language models / K. McGuffie, A. Newhouse. https://doi.org/10.48550/arXiv.2009.06807. Text: electronic.

12. *On the Dangers* of Stochastic Parrots: Can Language Models Be Too Big? / E. M. Bender [et al]. https://doi.org/10.1145/3442188.3445922. Text: electronic.

13. *Extracting* training data from large language models / N. Carlini [et al.]. Text: electronic // 30th USENIX Security Symposium (USENIX Security 21). 2021. P. 2633–2650. URL: https://www.usenix.org/system/files/sec21-carlini-extracting.pdf.

14. *On the Opportunities* and Risks of Foundation Models / R. Bommasani [et al.]. URL: https://arxiv.org/pdf/2108.07258.pdf. Text: electronic.

15. *Мониторинг* мнений: январь – февраль 2020. Текст: непосредственный // Мониторинг общественного мнения: экономические и социальные перемены. 2020. № 1 (155). С. 250–275. URL: https://monitoringjournal.ru/index.php/monitoring/article/view/1269.