

Захарова Оксана Игоревна,
к.т.н., доцент, зам. зав. НИЛ ИИ,
Поволжский государственный университет
телекоммуникаций и информатики
Zakharova Oksana Igorevna,
Candidate of Technical Sciences,
Associate Professor, Deputy Head. NEIL II,
Volga Region State University of
Telecommunications and Informatics

Воробьев Илья Романович, магистрант,
Поволжский государственный университет
телекоммуникаций и информатики
Vorobyov, Ilya Romanovich, Master's student,
Volga State University
of Telecommunications and Informatics

**ПРОГНОЗЫ НА БУДУЩЕЕ: ВОЗМОЖНЫЕ НАПРАВЛЕНИЯ РАЗВИТИЯ
МЕТОДОВ ОБРАБОТКИ ТЕКСТОВЫХ ДАННЫХ
FORECASTS FOR THE FUTURE: POSSIBLE DIRECTIONS FOR THE
DEVELOPMENT OF TEXT DATA PROCESSING METHODS**

Аннотация. Данная работа посвящена анализу перспективных направлений развития методов обработки текстовых данных. Рассматриваются такие ключевые тенденции, как эволюция больших языковых моделей, разработка энергоэффективных решений, интеграция мультимодальных данных. Выводы сделаны на основе анализа современных исследований и публикаций в области обработки естественного языка.

Abstract. This paper analyzes promising areas for the development of text data processing methods. Key trends such as the evolution of large-scale language models, the development of energy-efficient solutions, and the integration of multimodal data are considered. Conclusions are drawn from an analysis of modern research and publications in the field of natural language processing.

Ключевые слова: Обработка естественного языка, большие языковые модели, искусственный интеллект, машинное обучение, мультимодальные данные, энергоэффективность.

Keywords: Natural language processing, large language models, artificial intelligence, machine learning, multimodal data, energy efficiency.

Обработка текстовых данных, или обработка естественного языка (Natural Language Processing, NLP), является одной из наиболее динамично развивающихся областей искусственного интеллекта. Современные системы находят применение в самых разных сферах – от машинного перевода и голосовых помощников до анализа sentiment и автоматического реферирования текстов [1]. Развитие методов глубокого обучения, появление архитектур-трансформеров и накопление больших данных обусловили значительный прогресс в этой области [2]. Однако на пути дальнейшего развития стоят такие вызовы, как высокая ресурсоемкость, недостаточная интерпретируемость решений моделей, этические риски и необходимость работы в условиях ограниченных данных. Целью данного исследования является анализ ключевых направлений развития методов обработки текстовых данных, которые будут определять облик технологий в ближайшем будущем.



Большие языковые модели (БЯМ), такие как GPT, BERT и их аналоги, стали флагманом в области NLP. Их развитие движется в сторону увеличения объема параметров и обучающих данных, что позволяет достигать новых качественных уровней в понимании и генерации текста [3]. Однако трендом последнего времени становится не только масштабирование, но и повышение эффективности архитектур. Разрабатываются облегченные модели (например, DistilBERT, TinyBERT), сохраняющие высокое качество при значительном сокращении вычислительных затрат [4].

Важным направлением является адаптация БЯМ для узкоспециализированных предметных областей, таких как медицина, юриспруденция и финансы, где требуются точные знания и понимание специфической терминологии [5]. Персонализация моделей под нужды конкретного пользователя и развитие методов малосемплового обучения (few-shot learning) открывают возможности для их эффективного применения даже при ограниченных данных [6].

Современные системы все чаще переходят от работы исключительно с текстом к обработке мультимодальных данных, объединяющих текст, изображение, аудио и видео. Это позволяет создавать более комплексные и контекстуально aware-решения [7]. Например, мультимодальные модели способны генерировать описания изображений, отвечать на вопросы по видеоконтенту или создавать сценарии на основе визуальных образов.

Интеграция различных модальностей требует разработки новых архитектур и методов выравнивания представлений (representation alignment) из разнородных источников данных. Это направление напрямую связано с созданием более универсальных и гибких моделей искусственного интеллекта, способных воспринимать мир многогранно, подобно человеку.

Энергопотребление при обучении и эксплуатации крупных моделей стало серьезной проблемой, привлекающей внимание как с экономической, так и с экологической точек зрения [8]. В ответ на это развиваются методы оптимизации, включая квантование, дистилляцию знаний (knowledge distillation) и использование более эффективных архитектур внимания (attention mechanisms).

Параллельно растет запрос на интерпретируемость и объяснимость моделей NLP. Разрабатываются техники, позволяющие понять, на основании каких признаков и данных модель приняла то или иное решение (например, LIME, SHAP) [9]. Это критически важно для применения NLP в медицине, финансах и других ответственных областях, где необходимо доверие и контроль со стороны человека.

С расширением применения языковых моделей обострились вопросы их этического использования. Исследования направлены на снижение социальных предубеждений (bias), заложенных в обучающих данных, разработку методов фильтрации токсичного контента и обеспечение конфиденциальности [10]. Техники дифференциальной приватности (differential privacy) и федеративного обучения (federated learning) позволяют тренировать модели, не имея прямого доступа к персональным данным пользователей.

Проведенный анализ позволяет выделить несколько основополагающих направлений развития методов обработки текстовых данных. Во-первых, это оптимизация и специализация больших языковых моделей, делающая их более доступными и эффективными для конкретных задач. Во-вторых, интеграция текстовых методов с другими модальностями откроет путь к созданию по-настоящему многозадачных и контекстно-осознанных систем. В-третьих, неослабевающее внимание будет уделяться энергоэффективности, интерпретируемости и этической составляющей технологий ИИ. Совокупность этих тенденций свидетельствует о переходе от этапа бурного роста сложности моделей к этапу их качественного преобразования, направленного на практическую полезность, устойчивость и интеграцию в человеко-ориентированные системы.



Список литературы:

1. Девлин Дж. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding / Дж. Девлин, М.-В. Чанг, К. Ли, К. Туганова // Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. – 2019. – P. 4171–4186.
2. Attention is All You Need / А. Васивани [и др.] // Advances in Neural Information Processing Systems 30 (NIPS 2017). – 2017.
3. Language Models are Few-Shot Learners / Т.Б. Браун [и др.] // Advances in Neural Information Processing Systems 33 (NeurIPS 2020). – 2020. – P. 1877–1901.
4. Саньи К. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter / К. Саньи, Р. Дэбл, М. Мюллер [и др.] // arXiv preprint. – 2019. – arXiv:1910.01108.
5. BioBERT: a pre-trained biomedical language representation model for biomedical text mining / Дж. Лее [и др.] // Bioinformatics. – 2020. – Vol. 36, № 4. – P. 1234–1240.
6. K-Adapter: Infusing Knowledge into Pre-Trained Models with Adapters / Р. Ван [и др.] // arXiv preprint. – 2020. – arXiv:2002.01808.
7. ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks / Дж. Лу [и др.] // Advances in Neural Information Processing Systems 32 (NeurIPS 2019). – 2019.
8. Грин АИ / Р. Шварц [и др.] // Communications of the ACM. – 2020. – Vol. 63, № 12. – P. 54–63. – Пер. загл.: Green AI.
9. "Why Should I Trust You?": Explaining the Predictions of Any Classifier / М.Т. Рибейро, С. Сингх, К. Гестрин // Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. – 2016. – P. 1135–1144.
10. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? / Э. Бендакс [и др.] // Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21). – 2021. – P. 610–623.

