E

Проблемы искусственного интеллекта. 2026. N 1 (40). С. 150-158
Problems of Artificial Intelligence. 2026;1(40):150-158.
Системный анализ, управление и обработка информации, статистика
Научная статья

УДК 004.85
doi: 10.24412/2413-7383-2026-1-40-150-158

Н. В. Елисеева, В. Е. Петров
Московский государственный технологический университет «СТАНКИН»
127055, г. Москва, Вадковский переулок, д. 1

ТЕХНОЛОГИЯ СТИЛИСТИЧЕСКОЙ АДАПТАЦИИ НЕЙРОННОГО МАШИННОГО ПЕРЕВОДА

N. V. Eliseeva, V. E. Petrov
Moscow State University of Technology «STANKIN»
127055, Moscow, 1 Vadkovsky Lane

STYLISTIC ADAPTATION TECHNOLOGY OF NEURAL MACHINE TRANSLATION

Нейронный машинный перевод является ключевой технологией для автоматизации перевода текста. Крупные корпорации активно внедряют технологии нейронного машинного перевода в корпоративные процессы, используя их для глобальной коммуникации и локализации контента. В процессе внедрения и эксплуатации таких технологий у пользователей возникают новые требования к качеству машинного перевода, выходящие за рамки точной передачи содержания. Статья посвящена исследованию методов и технологий повышения качества нейронного машинного перевода. Предложенная в статье технология и результаты апробации показывают, что ее применение позволяет более точно передавать смысл и словарный запас оригинальных фраз.
Ключевые слова: нейронный машинный перевод, тональность текста, механизм внимания, сингулярная декомпозиция, оптимизация.

Neural machine translation is a key technology for automated text translation. Large corporations are actively integrating neural machine translation technologies into their corporate processes, using them for global communications and content localization. As these technologies are implemented and used, users face new demands on machine translation quality that go beyond the accurate rendering of content. This article explores methods and technologies for improving the quality of neural machine translation. The technology proposed in the article and the results of testing demonstrate that its use more accurately conveys the meaning and vocabulary of original phrases.
Key words: neural machine translation, attention mechanism, tonality of the text, singular value decomposition, optimization.

Введение

Актуальность исследования обусловлена тем, что существующие переводчики, демонстрируя высокую лексико-синтаксическую точность, слабо учитывают тональность и жанровые особенности текста. Практическая значимость стилистической адаптации перевода подтверждается активным развитием коммерческих сервисов, которые уже предлагают пользователям некоторые элементы настройки – например, выбор между формальным и неформальным стилем или региональным вариантом языка [1].

В эпоху глобализации и цифровой трансформации потребность в точном и быстром переводе текстов возрастает экспоненциально. Традиционные методы машинного перевода, такие как статистический машинный перевод и машинный перевод на основе правил, сталкиваются с ограничениями в передаче сложных семантических и различных нюансов языка [3-5]. Нейронный машинный перевод, основанный на искусственных нейронных сетях, предлагает качественно новый подход к решению этих задач. В основе нейронного машинного перевода лежит идея обучения модели преобразовывать последовательности слов исходного языка в последовательности слов целевого языка, учитывая сложные нелинейные зависимости и контекстуальные связи между ними.

Цель данного исследования – повысить качество нейронного машинного перевода на основе технологии стилистической адаптации.

Для достижения поставленной цели решены следующие задачи:

– разработана технология стилистической адаптации нейронного машинного перевода;
– разработана модель машинного нейронного перевода с управлением стилем и доменом перевода;
– проведено сравнение эффективности полученной модели с существующими моделями (без стилевой адаптации).

Анализ предметной области

Ключевые ограничения современных систем стилевого нейронного машинного перевода сводятся к трём взаимосвязанным аспектам (таблица 1) [6-13].

Во-первых, отсутствуют надёжные метрики стилевой эквивалентности. Стиль – многомерная величина (тон, регистр, эмоциональная окраска, идиолект), а потому единой автоматической меры пока нет. Разработчики вынуждены комбинировать косвенные показатели (BLEU, классификация стиля) с дорогой ручной экспертизой, что затрудняет воспроизводимое сравнение систем и замедляет прогресс [5].

Во-вторых, недостаточно параллельных корпусов с пометкой стиля. Немногие открытые наборы (GYAFC, Shakespeare → Modern Eng. и др.) покрывают лишь отдельные языки, жанры и дихотомии («формальный / неформальный» и т. п.). Большие монолингвальные коллекции тональности или формальности не решают задачу: без выровненных пар модель не получает прямого сигнала, как менять стиль, сохраняя смысл. В результате алгоритмы переобучаются на узких доменах или полагаются на синтетические данные, что снижает их надёжность [5].

Наконец, компромисс между сохранением стиля и семантической точностью. Сильный контроль стиля нередко ведёт к потере нюансов содержания, снижению беглости и появлению неестественных конструкций. Попытки развязать стиль и смысл (пост-обработка, управляющие теги) лишь частично смягчают проблему.

Таблица 1 - Основные проблемы текущих решений, их причины и последствия

Проблема	Причины	Последствия
Оценка стилевой эквивалентности	Многомерность понятия стиля; отсутствуют единые метрики и эталоны оценки; высокая субъективность человеческих суждений.	Затруднено сравнение систем; необходимость ручной оценки приводит к росту затрат и рисков несогласованности результатов.
Нехватка стилевых корпусов	Недостаток параллельных данных для разных стилей; ограниченность существующих датасетов (по языкам, жанрам); сложности сбора и разметки.	Ограниченная обучаемость моделей; узкая применимость; необходимость использования синтетических данных снижает достоверность.
Конфликт стиля и точности перевода	Стилевые признаки переплетены с содержанием на лингвистическом уровне; сильный контроль стиля приводит к изменениям лексики и синтаксиса.	Ухудшение смысловой точности при усилившемся стиле; появление неестественных или ошибочных конструкций; снижение общего качества и надёжности перевода.

Таким образом, для дальнейшего развития требуются: комплексная автоматическая оценка, учитывающая стиль, точность и естественность; масштабные многоязычные параллельные корпусы со стилевой аннотацией; архитектуры, минимизирующие конфликт «стиль – смысл». Решение этих задач снимет текущие ограничения и повысит практическую ценность стилевых настроек.

Постановка задачи

Для преодоления ограничений существующих систем нейронного машинного перевода предлагается интегрировать в нейронный машинный перевод комплексное решение, сочетающее динамическую адаптацию к домену, механизм избирательного внимания, расширенную языковую модель и оптимизацию матричных операций [14-20].

На входе система получает текст, который может включать разнообразные речевые конструкции, редкие слова, доменные термины, а иногда и размытый контекст. Вместо того, чтобы опираться только на стандартное «энкодер-декодер» преобразование или единственный метод понижения размерности, система учитывает несколько дополнительных компонентов, отвечающих за устойчивость и многофункциональность.

Модуль адаптации к предметной области использует предварительно собранные параллельные корпусы в заданном домене и автоматически подстраивается под терминологические и стилистические особенности.

Механизм избирательного внимания учитывает не только прямую связь между входным и выходным текстом, но и семантические перекрестные соответствия внутри самого исходного предложения.

Повышение точности перевода неизбежно порождает вопрос вычислительных затрат. Поэтому в предлагаемом подходе использован метод сингулярного разложения (SVD) для оптимизации матричных операций.

Адаптация модели к предметной области

Многие системы машинного перевода страдают от неспособности точно отразить редкую терминологию, свойственную специфическим доменам. В предлагаемой архитектуре модель не просто обучается на «общем» корпусе, а проходит дополнительную фазу дообучения на параллельных данных, относящихся к нужной предметной области. Данный процесс можно формально описать через функцию потерь, в которую вводится «доменная» компонента:

$$L_{total} = L_{NMT} + \beta L_{domain} \quad (1)$$

где L_{NMT} – стандартная кросс-энтропийная потеря, оценивающая расхождение между предсказанным и истинным переводами, а L_{domain}– штраф за неверное распознавание или передачу терминов и специфических выражений. Чтобы учесть лексические особенности, вводится набор векторных представлений dk для каждого домена (или тематического кластера). На этапе дообучения каждый входной пример «помечается» дополнительным признаком домена, и скрытые состояния корректируются согласно схеме:

$$h_i' = h_i + g\,(d_k, h_i) \quad (2)$$

где g – некий оператор (например, аффинное преобразование с нелинейной активацией), усиливающий или подавляющий отдельные измерения векторного представления h_i. Тем самым модель вникает в тонкости специализированного словаря, не теряя при этом общих навыков перевода.

Расширенный механизм внимания

Стандартный декодер обычно вычисляет коэффициенты внимания $\alpha_{t,i}$ на основании скалярного произведения или аддитивной функции оценки между скрытым состоянием декодера s_{t-1} и скрытым состоянием энкодера h_i. Это можно записать так:

$$e_{t,i} = score\,(s_{t-1}, h_i), \alpha_{t,i} = \frac{exp(e_{t,i})}{\sum_{k=1}^{T_x} exp(e_{t,k})} \quad (3)$$

Вместо одной функции $score$ предложено ввести дополнительное семантическое представление $\phi\,(h_i, h_i)$ для токенов внутри самого входного предложения.

Идея состоит в том, чтобы модель учитывала не только пару «декодер-энкодер», но и «взаимодействие внутри энкодера».

Функция $\phi\,(h_i, h_i)$ может быть реализована как косинусная похожесть или любая другая мера семантической близости. За счет этого декодер «подсвечивает» те фрагменты предложения, которые, с точки зрения внутренней лексико-семантической структуры, наиболее связаны друг с другом и потому потенциально важны для корректного перевода.

Оптимизация на основе сингулярного разложения

Для уменьшения вычислительной сложности вводится разложение матриц весов с использованием сингулярного разложения (SVD), что позволяет аппроксимировать большие матрицы меньшими ранга r, снижая тем самым количество вычислений.

К примеру, когда имеется большая матрица $W \in R^{mxn}$, вычисляющая промежуточное преобразование:

$$W = U \sum V^T \quad (4)$$

где U и V — ортонормированные матрицы, а $\sum$ — диагональная матрица сингулярных значений. Если оставляем только $r < \min(m, n)$ наибольших сингулярных чисел, то можно аппроксимировать W как произведение существенно меньших матриц U_r и V_r. Тем самым скорость вычислений возрастает, поскольку умножение на W заменяется на два умножения меньшей размерности.

При разумном выборе r качество перевода почти не страдает, зато модель обучается быстрее и требует меньше памяти. Дополнительно можно также применять квантование (сведение весов к числам с меньшим числом бит, например 8-битное или даже 4-битное представление), чтобы добиться еще большего выигрыша в скорости на графических ускорителях.

Алгоритм обучения модели

Обучение модели производилось путем минимизации функции потерь на основе кросс-энтропии:

$$L(\theta) = -\sum_{t=1}^{T_y} logP(y_t|y_{<t}, X; \theta) \quad (5)$$

Оптимизация параметров осуществлялась с использованием стохастического градиентного спуска и его адаптивных вариантов, таких как: Adam или RMSprop. Для обеспечения сходимости и предотвращения переобучения используются методы регуляризации:

- Dropout: случайное отключение нейронов с вероятностью p.
- L2-регуляризация: добавление штрафа за большие значения весов в функцию потерь [5], [10].

Экспериментальная оценка

Для оценки эффективности предлагаемого метода были проведены эксперименты на корпусе WMT 2014 English-German. Модель была реализована с использованием фреймворка PyTorch и обучалась на ядре GPU NVIDIA Tesla T4 с 16ГБ видеопамяти.

Параметры эксперимента:

- Размер словаря: 16 000 токенов, полученных с помощью Byte Pair Encoding (BPE);
- Размерность эмбеддингов и скрытых состояний: 256;
- Количество слоев энкодера и декодера: 4;
- Ранг аппроксимации r для метода сингулярного разложения SVD: 64.
- Метрики оценки:
  - BLEU: стандартная метрика качества перевода;
  - ROUGE: измерение совпадения фраз и структур между переводом и эталоном;

Таблица 2. Результаты эксперимента

Модель	BLEU	ROUGE	Время обучения (ч)
Базовый трансформер	22.8	13.3	36
Трансформер с предложенным авторским методом	24.1	14.2	28

Результаты показывают, что использование предложенной технологии приводит к улучшению качества перевода на 1.3 пункта BLEU и 0.9 пункта ROUGE по сравнению с базовой моделью. Сокращение времени обучения более чем на 20% свидетельствует об эффективности оптимизации вычислительной сложности.

Выводы

Предложенный комплексный подход и технология стилистической адаптации нейронного машинного перевода, улучшенного механизма внимания и оптимизации матричных операций демонстрируют повышение качества перевода и снижение вычислительной сложности, что делает модель более эффективной для практического применения.

Предлагаемое решение не ограничивается «усложнением» существующих методов машинного перевода одним инструментом (например, включением метода сингулярного разложения SVD). Оно многоуровневое: обладает усовершенствованным механизмом внимания, учитывающим дополнительные семантические связи в тексте, комплексной системой доменной адаптации, а также гибкой оптимизацией матричных операций под конкретную среду выполнения.

И если каждая из этих идей по отдельности встречается в современных исследованиях [8-10], [14-16], то их одновременное объединение в единую архитектуру создаёт принципиально новую парадигму.

Полученные в результате апробации показатели качества означают, что переводы ближе к эталону: модель с расширенным вниманием и доменной адаптацией точнее передает смысл и лексику исходных фраз. Хотя разница в метрических значениях кажется умеренной, в контексте машинного перевода даже улучшение на несколько баллов BLEU считается значительным достижением, особенно в таких сложных задачах, как перевод между разными языками или специализированными текстами.

При анализе ошибок было отмечено, что базовая модель чаще допускала неточности при переводе терминов и некоторых идиоматических выражений, в то время как предложенная автором версия более успешно справлялась с такими случаями, благодаря учёту контекстной семантики и адаптации к стилю предметной области.

Список литературы

1. Wang Y., Sun Z., Cheng S., Zheng W., Wang M. Controlling styles in neural machine translation with activation prompt // Findings of the Association for Computational Linguistics. 2023. P. 163–172.
2. Tan Z., Wang S., Yang Z., Chen G., Huang X., Sun M., Liu Y. Neural machine translation: a review of methods, resources, and tools. 2019. 112 p.
3. Servan C., Crego J., Senellart J. Domain specialization: a post-training domain adaptation for neural machine translation [Электронный ресурс]. arXiv:1612.06141, 2016. Режим доступа: https://arxiv.org/abs/1612.06141 (дата обращения: 11.10.2024).
4. Moslem Y., Way A., Haque R., Kelleher J. D. Domain-specific text generation for machine translation // AMTA-2022 – MT Research Track. 2022. P. 14–30.
5. Mukherjee S., Dusek O. Text style transfer: an introductory overview [Электронный ресурс]. arXiv:2407.14822, 2024. Режим доступа: https://arxiv.org/abs/2407.14822 (дата обращения: 22.01.2025)
6. Перевод и искусственный интеллект: проблемы и пути развития / С.А. Лосева. 2022.
7. Тагушева Н.Ю. (2015). Машинный перевод [Tagusheva N.Y. A Machine translation] // Материалы научно-практической конференции «Ломоносов 2015». Москва: МГУ.
8. Wang, X., Tu, Z., & Zhang, M. (2018). Incorporating statistical machine translation word knowledge into neural machine translation. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 26 (12), 2255–2266.
9. Kingma, D. P., and Ba, J. "Adam: A method for stochastic optimization." arXiv preprint arXiv:1412.6980 (2014).

10. Kudo, T., and Richardson, J. "SentencePiece: A simple and language-independent subword tokenizer and detokenizer for neural text processing." EMNLP (2018): 66-71.
11. Бондаренко, В. И. Анализ эффективности глубоких языковых моделей для задачи определения тональности русскоязычных текстов // В. И. Бондаренко, В. О. Елисеев, Т. В. Ермоленко // Проблемы искусственного интеллекта. - 2024. № 1 (32). - С. 51-62.
12. Харламов, А. Семантический анализ текста с использованием искусственных нейронных сетей на основе нейроподобных элементов с временным суммированием сигналов // А. Харламов, Е. Самаев, Д. Кузнецов и др. // Проблемы искусственного интеллекта. 2023. № 3 (30).
13. Ниценко А. В., Шелепов В. Ю. Об использовании семантической информации для снятия омонимии именительного и винительного падежа (как элемента создания онтологии) // Проблемы искусственного интеллекта. 2024. № 4 (34). С. 16-24
14. Zhang, Z., Liu, S., Li, M., Zhou, M., & Chen, E. (2018). Bidirectional generative adversarial networks for neural machine translation. In Proceedings of the 22nd Conference on Computational Natural Language Learning, pp. 190–199, Brussels, Belgium. Association for Computational Linguistics.
15. Stern, M., Chan, W., Kiros, J. R., & Uszkoreit, J. (2019). Insertion Transformer: Flexible sequence generation via insertion operations. arXiv preprint arXiv:1902.03249.
16. Stahlberg, F., Saunders, D., de Gispert, A., & Byrne, B. (2019). In Proceedings of the Fourth Conference on Machine Translation: Shared Task Papers. Association for Computational Linguistics.
17. Medina, J. R., & Kalita, J. (2018). Parallel attention mechanisms in neural machine translation. In 2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA), pp. 547–552.
18. McCandlish, S., Kaplan, J., Amodei, D., & Team, O. D. (2018). An empirical model of large-batch training. arXiv preprint arXiv:1812.06162.
19. Junczys-Dowmunt M., Grundkiewicz R., Dwojak T., Hoang H., Heafield K., Neckermann T., Seide F., Germann U., Aji A., Bogoychev N., Martins A., Birch A. Marian: fast neural machine translation in C++ // Proc. 56th Annual Meeting of the Association for Computational Linguistics: System Demonstrations. Melbourne: Association for Computational Linguistics, 2018. P. 116–121.
20. Булатов М.И., Елисеева Н.В., Петров В.Е. Технология повышения качества обучения искусственной нейронной сети в задачах управления объектами дорожно-транспортной инфраструктуры // International Journal of Open Information Technologies. 2024. Т. 12, № 4. С. 87-92

References

1. Wang Y., Sun Z., Cheng S., Zheng W., Wang M. Controlling styles in neural machine translation with activation prompt // Findings of the Association for Computational Linguistics. — 2023. — P. 163–172.
2. Tan Z., Wang S., Yang Z., Chen G., Huang X., Sun M., Liu Y. Neural machine translation: a review of methods, resources, and tools. 2019. 112 p.
3. Servan C., Crego J., Senellart J. Domain specialization: a post-training domain adaptation for neural machine translation — arXiv:1612.06141, 2016. URL: https://arxiv.org/abs/1612.06141
4. Moslem Y., Way A., Haque R., Kelleher J. D. Domain-specific text generation for machine translation // AMTA-2022 – MT Research Track. 2022. P. 14–30.
5. Mukherjee S., Dusek O. Text style transfer: an introductory overview — arXiv:2407.14822, 2024. Режим доступа: https://arxiv.org/abs/2407.14822
6. Loseva, S. A. (2022). Translation and Artificial Intelligence: Problems and Ways of Development.
7. Tagusheva, N. Y. (2015). Machine Translation. In Proceedings of the Scientific and Practical Conference "Lomonosov 2015". Moscow: Moscow State University.
8. Wang, X., Tu, Z., & Zhang, M. (2018). Incorporating statistical machine translation word knowledge into neural machine translation. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 26 (12), 2255–2266.
9. Kingma, D. P., and Ba, J. "Adam: A method for stochastic optimization." arXiv preprint arXiv:1412.6980 (2014).
10. Kudo, T., and Richardson, J. "SentencePiece: A simple and language-independent subword tokenizer and detokenizer for neural text processing." EMNLP (2018): 66-71.
11. Bondarenko, V. O. Analyzing the effectiveness of deep language models for the task of tone detection in russian-language texts // V. I. Bondarenko, V. O. Eliseev, T. V. Yermolenko // Problems of Artificial Intelligence. - 2024. № 1 (32). - С. 51-62.
12. Kharlamov A. Semantic text analysis using artificial neural networks based on neural-like elements with temporal signal summation // Kharlamov A., Samaev E., Kuznetsov D., Pantiukhin D. // Problems of Artificial Intelligence. - 2023. № 3 (30).

13. Nicenko A. V., Shelepov V. Ju. The use of semantic information to disambiguate the nominative/accusative homonyms: an element of creating ontology // Problems of Artificial Intelligence. 2024. № 4 (34). C. 16-24
14. Zhang, Z., Liu, S., Li, M., Zhou, M., & Chen, E. Bidirectional generative adversarial networks for neural machine translation. In Proceedings of the 22nd Conference on Computational Natural Language Learning, Brussels, Belgium, Association for Computational Linguistics. 2018. P. 190–199.
15. Stern, M., Chan, W., Kiros, J. R., & Uszkoreit, J. (2019). Insertion Transformer: Flexible sequence generation via insertion operations. arXiv preprint arXiv:1902.03249.
16. Stahlberg, F., Saunders, D., de Gispert, A., & Byrne, B. (2019). In Proceedings of the Fourth Conference on Machine Translation: Shared Task Papers. Association for Computational Linguistics.
17. Medina, J. R., & Kalita, J. (2018). Parallel attention mechanisms in neural machine translation. In 2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA), pp. 547–552.
18. McCandlish, S., Kaplan, J., Amodei, D., & Team, O. D. (2018). An empirical model of large-batch training. arXiv preprint arXiv:1812.06162.
19. Junczys-Dowmunt M., Grundkiewicz R., Dwojak T., Hoang H., Heafield K., Neckermann T., Seide F., Germann U., Aji A., Bogoychev N., Martins A., Birch A. Marian: fast neural machine translation in C++ // Proc. 56th Annual Meeting of the Association for Computational Linguistics: System Demonstrations. — Melbourne: Association for Computational Linguistics, 2018. — P. 116–121.
20. M.I. Bulatov, N.V. Eliseeva, V.E. Petrov Technology for improving the quality of training of an artificial neural network in problems of managing road transport infrastructure objects // International Journal of Open Information Technologies. - 2024. - Vol. 12, No. 4. - P. 87-92

RESUME

N. V. Eliseeva, V. E. Petrov

Stylistic adaptation technology of neural machine translation

The main goal of the paper is to improve the quality of technical translation in production and corporate communication. The results show that this method, more accurately conveys the meaning and vocabulary of the original phrases.

In the era of globalization and digital transformation, the demand for precise and rapid text translation is increasing exponentially. Traditional machine translation methods, such as statistical machine translation (SMT) and rule-based machine translation (RBMT), encounter limitations in conveying complex semantic structures and various linguistic nuances.

The proposed solution is not limited to «tightening» existingmachine translation methods using a single tool (for example, the inclusion of SVD). It is multi-layered: it has an improvedattention mechanism that takes into account additional semantic connections-within the text, a comprehensive system of domain adaptation, as well as flexibleoptimi-zation of matrix operations for a specific execution environment. And if each of these ideas can be found individually inmodern research, then their simultaneous unification into a single architecture creates a fundamentally new paradigm.

The results indicate that the use of the proposed method leads to an improvement in translation quality by 1.3 BLEU points and 0.9 ROUGE points compared to the baseline model. Additionally, the reduction in training time by more than 20% demonstrates the effectiveness of optimizing computational complexity.

This study presents a method for enhancing the attention mechanism in neural machine translators by integrating semantic information and optimizing matrix operations. Experimental results demonstrate an improvement in translation quality and a reduction in computational complexity, thereby making the model more efficient for practical applications.

РЕЗЮМЕ

Н.В. Елисеева, В. Е. Петров

Технология стилистической адаптации нейронного машинного перевода

Основная цель статьи – повышение качества технического перевода в сфере производственной и корпоративной коммуникации. Результаты показывают, что данный метод точнее передает смысл и лексику исходных фраз.

В эпоху глобализации и цифровой трансформации спрос на точный и быстрый перевод текстов растет экспоненциально. Традиционные методы машинного перевода, такие как статистический машинный перевод и машинный перевод на основе правил, сталкиваются с ограничениями при передаче сложных семантических структур и различных языковых нюансов.

Предлагаемое решение не ограничивается «усилением» существующих методов машинного перевода с помощью одного инструмента (например, включением SVD). Он многослойный: обладает усовершенствованным механизмом внимания, учитывающим дополнительные семантические связи в тексте, комплексной системой адаптации к домену, а также гибкой оптимизацией матричных операций под конкретную среду выполнения. И если каждая из этих идей по отдельности встречается в современных исследованиях, то их одновременное объединение в единую архитектуру создаёт принципиально новую парадигму.

Результаты показывают, что использование предлагаемого метода приводит к улучшению качества перевода на 1,3 балла BLEU и 0,9 балла ROUGE по сравнению с базовой моделью. Кроме того, сокращение времени обучения более чем на 20% демонстрирует эффективность оптимизации вычислительной сложности.

В данном исследовании представлен метод улучшения механизма внимания в нейронных машинных переводчиках путем интеграции семантической информации и оптимизации матричных операций. Экспериментальные результаты демонстрируют улучшение качества перевода и снижение вычислительной сложности, что делает модель более эффективной для практического применения.

Елисеева Наталья Владимировна – доцент, ФГБОУ ВО "МГТУ "СТАНКИН", 127055, Москва, Вадковский пер., д. 1, телефон +7(909) 168-7362, n.eliseeva@stankin.ru. *Область научных интересов:* нейронные сети, экспертные системы, онтологии. ORCID 0009-0004-1240-4552

Петров Валерий Евгеньевич – доцент, ФГБОУ ВО "МГТУ "СТАНКИН", 127055, Москва, Вадковский пер., д. 1, телефон +7(916) 166-7105, v.petrov@stankin.ru. *Область научных интересов:* математическое моделирование систем искусственного интеллекта, нейронные сети, экспертные системы. ORCID 0009-0006-9112-8182

Eliseeva Nataly Vladimirovna – associate professor of "MSUT "STANKIN", 127055, Moscow, 1 Vadkovsky Lane, phone +7(909) 168-7362, n.eliseeva@stankin.ru. Research interests: neural networks, expert systems, ontology. ORCID 0009-0004-1240-4552

Petrov Valeriy Evgenivech – associate professor of "MSUT "STANKIN", 127055, Moscow, 1 Vadkovsky Lane, phone +7(916) 166-7105, v.petrov@stankin.ru. Research interests: mathematical modeling of artificial intelligence systems, neural networks, expert systems. ORCID 0009-0006-9112-8182

Статья поступила в редакцию 03.12.2025.