«УТВЕРЖДАЮ»
Директор Федерального государственного учреждения «Федеральный исследовательский центр «Информатика и управление» Российской академии наук»,

_____ М.А. Посыпкин

«26» 05 2026 г.

ЗАКЛЮЧЕНИЕ

Федерального государственного учреждения «Федеральный исследовательский центр «Информатика и управление» Российской академии наук

Диссертация Скачкова Николая Андреевича на тему: «Вероятностные нейросетевые модели понимания и генерации текстов естественного языка» выполнена в отделе №14 ФИЦ ИУ РАН .

Скачков Николай Андреевич, 1997 года рождения, гражданин России, в 2021 году окончил Федеральное государственное бюджетное образовательное учреждение высшего образования «Московский государственный университет имени М.В. Ломоносова» по направлению 01.04.02 «Прикладная математика и информатика».

Скачков Николай Андреевич окончил аспирантуру очной формы обучения в ФИЦ ИУ РАН по научной специальности 05.13.17 «Теоретические основы информатики», дата окончания 22 сентября 2025 года.

Тема диссертационного исследования утверждена на заседании ученого совета, протокол заседания Ученого совета ФИЦ ИУ РАН № 1 от 21 января 2022г.

В период подготовки диссертации соискатель ученой степени Скачков Николай Андреевич работал в ООО «Яндекс.Технологии» в подразделении «Служба качества МТ» с 2018 г. по 2023 г. в должности разработчика.

Справка о периоде обучения в период (с 23.09.2021 г. по 22.09.2025 г.) с результатами сдачи кандидатских экзаменов по специальности 2.3.8 «Информатика и информационные процессы» выдана 08.04.2026 г. ФИЦ ИУ РАН.

Научный руководитель д.ф-м.н. проф. Воронцов Константин Вячеславович работает в МГУ в должности зав. кафедрой математических методов прогнозирования факультета вычислительной математики и кибернетики ФГБОУ ВО «Московский государственный университет имени М.В.Ломоносова».

По итогам обсуждения принято следующее заключение.

Актуальность темы

Актуальность исследования обусловлена стремительным ростом объёмов многоязычного контента и цифровизацией коммуникаций, что делает машинный перевод критически важной технологией для науки, бизнеса и государственного сектора. Несмотря на революцию, связанную с появлением архитектуры Transformer и больших языковых моделей, качество перевода остаётся ограниченным системными факторами: дефицитом параллельных данных для малоресурсных языков и узких доменов, нехватка контекстной

информации при переводе фрагментов документов, накоплением ошибок в авторегрессионной генерации, а также низким качеством обучающих данных. Современные системы демонстрируют типовые сбои — недостаточные и избыточные переводы, искажения смысла, — которые плохо устраняются стандартной функцией потерь максимизации правдоподобия.

В этих условиях особую значимость приобретают методы, которые решают обозначенные проблемы и позволяют улучшить качество перевода в различных условиях.

В диссертационной работе предлагаются новые вероятностные, архитектурные и обучающие методы генеративных моделей машинного перевода, ориентированных на преодоление проблем и ограничений, возникающих при создании современных систем автоматического машинного перевода. Описаны способы преодоления следующих существенных ограничений:

1) Недостаток параллельных данных, возникающий при обучении моделей перевода, например, на более редких языках.

2) Систематические ошибки моделей перевода, возникающие из-за некачественных обучающих данных, собранных с помощью эвристических алгоритмов.

3) Систематические ошибки моделей перевода, возникающие из-за недостатков функции потерь, используемой обычно при обучении моделей перевода.

4) Контекстные ошибки перевода, возникающие при независимом переводе предложений или фрагментов текста в задаче перевода документов.

Личное участие соискателя ученой степени в получении результатов, изложенных в диссертации

Все результаты, представленные в диссертационном исследовании, получены лично соискателем под научным руководством К.В.Воронцова.

Степень достоверности результатов проведенных соискателем ученой степени исследований

Достоверность полученных результатов подтверждена экспериментальной проверкой предлагаемых методов на реальных данных; публикациями результатов исследования в рецензируемых научных изданиях и конференциях по машинному обучению; воспроизводимостью результатов исследования при использовании различных тестовых наборов данных из открытых источников.

Научная новизна работы

Научная новизна работы заключается в новых методах обучения перевода, таких как совместное обучение прямой и обратной моделей, переупорядочивание гипотез на основе экспертной разметки, метод маскирования входа при обучении моделей перевода, использование тематической сегментации в задаче перевода с контекстом. Данные методы имеют теоретическое обоснование и показана практическая эффективность их применения для улучшения качества перевода, что подтверждено ростом автоматических метрик оценки качества перевода и экспертной разметкой.

Таким образом, тема диссертации отвечает ключевым проблемам текущего этапа развития машинного перевода: повышает качество в условиях ограниченных данных, улучшает согласованность и надёжность генерации, расширяет применимость к длинным

документам и доменно-специфичным сценариям и органично сочетается с современным направлением адаптации больших языковых моделей под задачу перевода.

Теоретическая значимость

1. Предложен новый вероятностный метод совместного обучения прямой и обратной моделей перевода, использующий функцию потерь максимизации правдоподобия циклического перевода. Выведена формула градиента для градиентных методов оптимизации.

2. Предложена теоретическая модель маскирования, являющаяся обобщением для моделей постредактирования, перевода и маскированного языкового моделирования.

3. Теоретически обоснован метод постобработки E-шага для построения сегментирующих тематических моделей. Выведена формула регуляризатора на основе требования разреженности тематик внутри текстовых фрагментов.

Практическая значимость результатов проведенных соискателем ученой степени исследований

Практическая значимость заключается в разработке методов, повышающих качество нейросетевого машинного перевода как для крупных языковых пар, так и для малоресурсных направлений, где классические подходы часто не работают.

1. Предложенный метод совместного обучения с использованием обратной модели позволяет заметно улучшать точность при дефиците параллельных данных и может быть напрямую внедрён в промышленные системы, расширяя их на новые языки и домены без существенных затрат на сбор обучающих данных.

2. Для преодоления типовых систематических ошибок (недостаточных и избыточных переводов) предложено переупорядочивание гипотез по человеческим предпочтениям, что даёт эффективную доменную адаптацию даже без эталонных параллельных текстов, снижает объём ручной правки и приближает качество к профессиональному уровню.

3. Предложенная функция потерь маскировки входа повышает лингвистическую и грамматическую корректность перевода и позволяют использовать одноязычные корпуса, что удешевляет внедрение.

4. Для длинных и сложноструктурированных документов предложена тематическая сегментация на основе тематического моделирования, которая разбивает тексты на логически цельные фрагменты, повышая связность контекстного перевода.

Апробация работы

Основные результаты работы докладывались на:

1. XXVIII Международная конференция студентов, аспирантов и молодых учёных «Ломоносов», Москва, 2021 г.

2. 20-я Всероссийская конференция с международным участием «Математические методы распознавания образов» (ММРО-2021), Москва, 2021 г.

3. Международная конференция «Интеллектуализация обработки информации», г. Москва, 2022

4. Международная конференция «Интеллектуализация обработки информации 2024», 2024.

Также результаты работы докладывались на научных семинарах ФИЦ ИУ РАН и ВМК МГУ.

Полнота изложения материалов диссертации в публикациях

Основные результаты диссертационного исследования опубликованы в 6 работах.

Публикации соответствуют теме диссертационного исследования и полностью раскрывают её основные положения.

Ценность научных работ соискателя ученой степени

Практическая ценность результатов исследования заключается в разработке новых методов, направленных на повышение качества нейросетевого машинного перевода — как для крупных языковых пар с большими корпусами данных, так и для малоресурсных направлений, где традиционные методы оказываются неэффективны.

1) Метод совместного обучения прямой и обратной модели перевода улучшает качества перевода на 0.5 BLEU на малоресурсном англо-финском направлении перевода.

2) Метод переупорядочивания гипотез на основе человеческой разметки улучшает качество перевода на русско-английском направлении на 0.6 BLEU, а также уменьшает долю недостаточных переводов с 4% до 1%. Результы разметки переводов данного метода показывают, что люди на 15% чаще выбирают перевод, полученный с помощью предложенного подхода. Данный подход удешевляет адаптацию моделей машинного перевода к специализированным областям с нехваткой параллельных данных.

3) Метод маскировки входа улучшает качество перевода на 0.4 BLEU в задаче перевода с английского на русский. Результы разметки переводов данного метода показывают, что люди на 5% чаще выбирают перевод, полученный с помощью предложенного подхода. Данный подход удешевляет создание автоматических моделей постредактирования переводов.

4) Метод сегментации с помощью тематических моделей улучшает качество контекстного перевода на 0.2 BLEU в задаче перевода с ангдийского на русский. Сам метод сегментации превосходит классические тематические модели на 5-10% по метрикам WindowDiff и Pk.

Публикации соискателя по теме диссертации

1. Skachkov N.A., Vorontsov K.V. Improving Topic Models with Segmental Structure of Texts // Computational Linguistics and Intellectual Technologies: Proc. Intern. Conf. "Dialogue 2018". 2018. Vol. 17. P. 652-661. (**Scopus)**

2. Shtekh G., Kazakova P., Nikitinsky N., Skachkov N. Applying Topic Segmentation to Document-Level Information Retrieval // Proceedings of the 14th Central and Eastern European Software Engineering Conference Russia. 2018. No 6. P. 1-6. (**Scopus)**

3. Скачков Н.А., Воронцов К.В. Улучшение качества машинного перевода с использованием обратной модели. // Автоматика и телемеханика. 2022. № 12, 31–43 **(ВАК, RSCI)** (*Перевод: Skachkov N.A., Vorontsov K.V. Improving the Quality of Machine Translation Using the Reverse Model // Automation and Remote Control. 2022. Vol. 83. P. 1897-1907.* (**Scopus**))

4. Воронцов К.В., Скачков Н.А. Упорядочивание гипотез в моделях перевода с использованием человеческой разметки // Известия Российской Академии Наук. Теория и системы управления. 2024. № 4. С. 121-128. **(ВАК, RSCI)** (*Перевод: Vorontsov K.V., Skachkov N.A. Hypotheses Re-ranking in Translation Models Using Human Markup // J. Comput. Syst. Sci. Int. 2024. V.63. N4. P.679–686.* (**WoS, Scopus**))

5. Elshin D., Karpachev N., Gruzdev B., Golovanov I., Ivanov G., Antonov A., Skachkov N., Latypova E., Layner V., Enikeeva E., Popov D., Chekashev A., Negodin V., Frantsuzova V., Chernyshev A., Denisov K. From General LLM to Translation: How We Dramatically Improve Translation Quality Using Human Evaluation Data for LLM Finetuning. // Proc. 9th Conf. Machine Translation. 2024. P. 247-252. (**Scopus**)

6. Skachkov N.A. Method of Input Masking for Training Translation Models. // Pattern Recognit. Image Anal. 2025. V. 35, P. 493–500. **(Scopus)**

Диссертация Скачкова Николая Андреевича на тему: «Вероятностные нейросетевые модели понимания и генерации текстов естественного языка» – законченная научно-квалификационная работа, которая соответствует: требованиям пунктов 9, 10, 14 Положения о присуждении ученых степеней, утвержденного постановлением Правительства Российской Федерации от 24 сентября 2013 г. № 842, Паспорту научной специальности 2.3.8 «Информатика и информационные процессы», технические науки, в частности пунктам:

4. Разработка методов и технологий цифровой обработки аудиовизуальной информации с целью обнаружения закономерностей в данных, включая обработку текстовых и иных изображений, видео контента. Разработка методов и моделей распознавания, понимания и синтеза речи, принципов и методов извлечения требуемой информации из текстов.
5. Лингвистическое обеспечение информационных систем и процессов. Методы и средства проектирования словарей данных, словарей индексирования и поиска информации, тезаурусов и иных лексических комплексов. Методы семантического, синтаксического и прагматического анализа текстовой информации для представления в базах данных и организации интерфейсов информационных систем с пользователями.
16. Автоматизированные информационные системы, ресурсы и технологии по областям применения (научные, технические, экономические, образовательные, гуманитарные сферы деятельности), форматам обрабатываемой, хранимой информации. Системы принятия групповых решений, системы проектирования объектов и процессов, экспертные системы и др.

Диссертация Скачкова Николая Андреевича на тему: «Вероятностные нейросетевые модели понимания и генерации текстов естественного языка» рекомендуется к защите на соискание ученой степени кандидата технических наук наук по специальности 2.3.8 «Информатика и информационные процессы»

Заключение принято на заседании отдела №14 ФИЦ ИУ РАН «30» апреля 2026 г., протокол № 1.

Присутствовало на заседании 10 человек.

Результаты голосования: «за» – 10, «против» – 0, «воздержалось» – 0.

Председатель заседания:
к.ф.-м.н.
в.н.с. отдела №14 И.Ю.Торшин

Секретарь:
инж-иссл. отдела №14 А.Н.Громов