На правах рукописи

Скачков Николай Андреевич

Вероятностные нейросетевые модели понимания и генерации текстов естественного языка

Специальность 2.3.8 —
«Информатика и информационные процессы»

Автореферат
диссертации на соискание учёной степени
кандидата технических наук

Москва — 2026

Работа выполнена в Федеральном исследовательском центре «Информатика и управление» Российской академии наук.

Научный руководитель: доктор физико-математических наук, профессор РАН
Воронцов Константин Вячеславович

Официальные оппоненты: **Максимов Николай Вениаминович,**
доктор технических наук, профессор,
Федеральное государственное автономное образовательное учреждение высшего образования «Национальный исследовательский ядерный университет «МИФИ»,
профессор кафедры «Финансовый мониторинг»

Ильвовский Дмитрий Алексеевич,
кандидат технических наук,
Национальный исследовательский университет «Высшая школа экономики»,
научный сотрудник международной лаборатории интеллектуальных систем и структурного анализа

Ведущая организация: Федеральное государственное бюджетное учреждение науки Всероссийский институт научной и технической информации Российской академии наук

Защита состоится «_____» _____2026 года на заседании диссертационного совета 24.1.224.03 при Федеральном исследовательском центре «Информатика и управление» Российской академии наук по адресу: 19333, Россия, Москва, ул. Вавилова, д. 42.

С диссертацией можно ознакомиться в библиотеке Федерального исследовательского центра «Информатика и управление» Российской академии наук и на сайте http://www.frccsc.ru/.

Автореферат разослан «_____» _____2026 года.

Ученый секретарь
диссертационного совета
24.1.224.03,
кандидат технических наук Рейер Иван Александрович

Общая характеристика работы

Актуальность темы. В последние десятилетия компьютерная лингвистика переживает революционные изменения благодаря стремительному развитию машинного обучения, искусственного интеллекта и вычислительных мощностей. Автоматический машинный перевод — одно из важнейших прикладных направлений искусственного интеллекта. Качественные системы машинного перевода (МП) становятся ключевым компонентом цифровой трансформации мирового информационного пространства, поддерживают межкультурную коммуникацию, трансграничную торговлю, международное сотрудничество в науке и технологиях.

Современные онлайн-переводчики позволяют мгновенно переводить тексты и устную речь между сотнями языков мира, а передовые системы, такие как Google Translate, DeepL или Яндекс Переводчик, обрабатывают ежедневно гигантские объемы информации. Вместе с тем, несмотря на впечатляющие качественные успехи в задаче машинного перевода, которые произошли в последние годы, существует ряд фундаментальных научных и технологических вызовов, стоящих на пути окончательного решения этой задачи. Так, переводы, выдаваемые даже самыми лучшими системами, нередко содержат неточности, грамматические и стилистические ошибки, искажения или потери смысла. Все эти ошибки заметно влияют на восприятие переведенного текста людьми. Данные проблемы приводят к тому, что итоговый текст выглядит неестественным с точки зрения языка и может создавать эффект «зловещей долины». Также ошибки в переводе напрямую мешают задаче пользователя — понять текст на иностранном языке и донести информацию до собеседника без искажения.

Особенно остро проблемы с качеством перевода могут проявляться при переводе сложных текстов с узкой специализированной лексикой, а также при переводе длинных текстов и при переводе между редкими языковыми парами, где есть дефицит параллельных данных. Иллюстрацией последней проблемы являются языки с относительно небольшим количеством носителей.

Стремление к переводу «человеческого уровня» — одна из больших нерешенных задач когнитивной вычислительной лингвистики. Решение этой большой задачи неразрывно связано с прорывами в близких областях, таких как интерпретируемость нейронных моделей, обобщаемость моделей на новые домены, обеспечение точности и полноты передачи смысловых элементов и передаваемой информации. Все это важнейшие исследовательские вопросы, затрагивающие вопросы безопасности и доверия к технологиям ИИ.

Разработку новых методов совершенствования генеративных моделей машинного перевода, повышающих достоверность, интерпретируемость и соответствие человеческим ожиданиям, можно без преувеличения отнести к числу приоритетных научных задач современного этапа развития искусственного интеллекта.

Все качественные проблемы моделей перевода можно разделить на классы для удобства интерпретации и идентификации проблем. В данной работе предлагается два способа классификации. Первый способ — это разделение ошибок по сути самой ошибки. С точки зрения такого подхода, можно выделить следующие типы систематических проблем современных моделей перевода:

- Ошибки недостаточного или избыточного перевода. В этом случае при переводе модель удаляет какую-то информацию, содержащуюся в исходном тексте или добавляет что-то, чего в изначальном тексте не было. Проблема такого рода может приводить к недостоверности перевода. При наличии даже небольшой ошибки общий смысл переводимого текста может заметно отличаться от того, что было сказано в оригинале.
- Ошибки неточного перевода. В этом случае при переводе модель меняет часть текста на отличающийся по смыслу. Например, меняет существительное на антоним. Эта грубая ошибка перевода, как и в предыдущем случае, приводит к значительной недостоверности переведенного текста.
- Ошибки гладкости. В этом случае при переводе модель не меняет смысла исходного текста, но перевод нарушает лингвистические нормы целевого языка. Например, в переводе появляются языковые конструкции, которые звучат неестественно с точки зрения носителя целевого языка. Такие переводы могут оставаться достоверными, но могут приводить к недопониманию и снижают доверие к модели машинного перевода.
- Ошибки контекста. Такого рода ошибки могут возникать при переводе достаточно длинных текстов и сводятся к тому, что модель неконсистентно переводит связанные понятия. Примером такой ошибки может быть отличающийся перевод именованной сущности, встречающейся в разных предложениях. Такие ошибки могут существенно влиять на понятность текста после перевода. К этим же ошибкам можно отнести проблемы сохранения структуры текста при переводе. Например, неправильный перевод списков и внутритекстовых сносок, а также других элементов форматирования.

Второй подход к описанию ошибок моделей машинного перевода заключается в разделении по причине возникновения систематических ошибок. Такой подход помогает диагностировать слабые места в общепринятых процессах построения систем перевода и понять способы преодоления возникающих проблем с качеством. Среди причин низкого качества систем перевода можно выделить следующие классы:

- Нехватка или низкое качество обучающих данных. Эта проблема остро возникает при обучении моделей перевода между языками, в которых хотя бы один язык является редким и плохо представленным в интернете. Для таких языковых пар затруднительно собрать обучающие корпуса объема, достаточного для получения высокого качества. Современные модели перевода имеют большое количество параметров и

требуют параллельных корпусов, состоящих из сотен тысяч переведенных фрагментов текста. Также данная проблема возникает при сборе обучающих корпусов по узким тематикам. Не все области деятельности человека хорошо представлены в интернете и других доступных источниках, что может приводить к низкому качеству перевода текстов со специальной лексикой.

– Ограничения вероятностной модели. Современные модели перевода зачастую учатся повторять увиденные во время обучения тексты слева направо. Вероятностные модели, приводящие к такой постановке задачи, имеют ряд ограничений и могут стимулировать модель совершать различные систематические ошибки.
– Контекстные ограничения. Исторически задача перевода решалась на уровне предложений и большинство обучающих корпусов и разработанных процессов их построения сильно завязаны на это ограничение. Из-за этого адаптация моделей к переводу более длинных фрагментов текста может быть проблематичной и контекстная согласованность переводов может заметно страдать.

Итак, комплексная проблема современного машинного перевода состоит в необходимости создания моделей и методов, которые бы обеспечивали выполнение сразу многих критериев. В первую очередь, модель должна обладать высоким качеством, то есть сохранять при переводе исходный смысл, не допуская избыточных, недостаточных и неточных переводов. Также переводы модели должны быть лингвистически приемлемыми с точки зрения целевого языка, в них не должны быть нарушены контекстные связи, представленные в исходном тексте, а также должна сохраняться структура исходного текста, что особенно актуально при переводе текстов в различных форматах, таких как HTML и субтитры.

Фундаментальные основы и методы нейросетевого машинного перевода, обработки естественного языка и вероятностного моделирования заложены в трудах отечественных и зарубежных ученых. Значительный прогресс в области был достигнут за счет развития архитектур последовательных моделей с механизмом внимания и трансформеров. Далее архитектурные изменения моделей и методов перевода шли в основном по пути решения конкретных прикладных проблем. Так, крупным направлением исследований являются подходы по улучшению качества и устойчивости переводов редких слов и языковых конструкций. Другим ярким примером является направление исследований неавторегрессионной генерации переводов. Эти методы позволяют при некотором ухудшении качества перевода добиться значительного прогресса в скорости перевода.

Несмотря на появление мощных архитектур и развитие методов обучения, проблема обучения качественных моделей в условиях нехватки параллельных данных остается актуальной. Существенный прогресс в решении этой задачи был достигнут за счет интеграции обратной модели перевода в обучение методом генерации синтетических обратных переводов. Обратная модель перевода

— это модель перевода, соответствующая переводу с целевого языка на входной. То есть для модели перевода с русского на английский язык обратной будет являться модель перевода с английского на русский язык. Использование синтетических обратных переводов позволяет на практике частично компенсировать нехватку параллельных данных в обучении за счет дистилляции обратной модели перевода. Кроме того, активно развивались методы интеграции обратной молели перевода в процессе обучения прямой модели. Тем не менее, такие подходы оказались труднопримениимыми на практике по мере прогресса в области увеличения размеров переводных моделей, так как при этом хранение в памяти вычислительного устройства одновременно прямой и обратной моделей становится менее эффективным, а сам процесс обучения оказывается менее устойчивым.

Отдельный интерес представляют из себя задачи по борьбе с систематическими ошибками нейронных моделей, такими как пропуски слов и избыточная генерация. Видов и даже классификаций систематических ошибок переводов на практике достаточно много, и каждый подход требует отдельной обработки каждого конкретного типа: либо с помощью модификации процесса обучения под этот тип ошибки, либо за счет контрастного обучения с синтетическими негативными примерами. Однако на практике негативные примеры формируются искусственно на основе жестких эвристических шаблонов, что делает невозможным охватить реальное многообразие ошибок модели. В то же время, несмотря на активное развитие нейронных метрик оценки качества перевода, обученных на разметках, выполненных человеком, методы прямой интеграции человеческих предпочтений непосредственно в процесс дообучения моделей перевода остаются недостаточно изученными.

В смежной области языкового моделирования значительный прогресс был достигнут за счет применения методов маскировки входа, представленных в моделях BERT, BART, mT5 и других. За счет изменения процесса обучения и усложнения функции потерь, используемой при обучении, в них удалось добиться более высокого качества на прикладных задачах понимания и генерации текстов. Однако эти подходы ориентированы преимущественно на задачи понимания текста или моноязычной генерации. Адаптация маскированных функций потерь для задачи перевода и их влияния на качество генераций требует более глубокого исследования.

Еще одной значимой областью является исследование проблем качества контекстного перевода. Многие системы перевода исторически делят входной текст на небольшие фрагменты: предложения или несколько предложений. При неудачном разделении входного текста могут теряться тематическая связанность и контекстная информация, необходимая, например, для разрешения анафор. Задача сегментации текста может решаться с помощью тематических моделей, которые по построению должны предоставлять качественные семантические представления слов и документов. Однако модификация тематических моделей

под эти задачи приводит к значительному усложнению математического вывода. Разработка методов восстановления сегментной структуры документа на базе аддитивной регуляризации без усложнения EM-алгоритма, а также применение этих методов для повышения качества перевода длинных документов изучены недостаточно.

Анализ описанной темы позволяет сформулировать основную научную проблему исследования. Наблюдается противоречие между потребностью общества в системах машинного перевода «человеческого уровня», отличающихся переводами с высокой смысловой точностью, лингвистической грамотностью и контекстной связностью для любых языков и документов любой длины, и ограничениями методов обучения, не дающих нужное качество моделей перевода.

Цель диссертационной работы — разработка новых вероятностных и архитектурных методов обучения генеративных моделей машинного перевода, ориентированных на преодоление обозначенных проблем, связанных с нехваткой обучающих данных, их недостаточным качеством, ограничениями стандартной авторегрессионной постановкой задачи перевода, а также некачественной сегментацией входного текста в задаче контекстного перевода. Эффективность предложенных методов оценивается с точки зрения общепринятых метрик качества в задаче машинного перевода, среди которых есть BLEU, BLEURT, COMET, а также разметка переводов на качество, выполненная человеком.

Для достижения поставленной цели необходимо было решить следующие **задачи**:

1. Разработать и теоретически обосновать новый метод совместного обучения с использованием обратной модели в условиях недостаточного объёма параллельных данных. Для этого необходимо было провести вероятностный вывод, определить оптимизируемую функцию потерь и эмпирически доказать, что интеграция обратной модели позволяет добавить дополнительную информацию в процесс обучения, что приводит к повышению качества и точности перевода для малоресурсных языковых направлений.
2. Предложить и реализовать метод переупорядочивания переводческих гипотез на основе разметки, выполненной человеком, для борьбы с систематическими ошибками перевода. В рамках этой задачи требовалось организовать эффективный процесс сбора разметки, внедрить обучение моделей на основе полученного ранжирования и продемонстрировать, что такой подход не только способствует снижению доли ошибок, но и повышает эффективность и экономичность доменной адаптации переводческих систем к домену электронной коммерции.
3. Разработать метод маскировки входа для моделей машинного перевода, нацеленный на уменьшение числа систематических ошибок грамотности и точности автоматических переводов. Было необходимо вывести новую вероятностную модель, реализовать её в качестве обучающей

функции потерь и провести серию экспериментов для оценки эффективности данного подхода в части моделирования сложных контекстных зависимостей при переводе.

4. Создать и экспериментально обосновать способ сегментации длинных текстов на основе тематического моделирования. Было необходимо показать эффективность использования тематических моделей в задаче сегментации. Для этого была построена тематическая модель и был реализован алгоритм автоматического определения границ сегментов с учётом семантической однородности. Далее эффективность предложенного метода сегментации необходимо было проверить на задаче документного перевода.

Научная новизна:

1. В работе впервые теоретически обоснован способ совместного обучения прямой и обратной модели, выведенный из решения задачи максимизации циклического правдоподобия, что позволяет выстроить процедуру совместного обучения прямой и обратной моделей без использования дополнительных языковых моделей или множества эвристик. В работе предложен более легкий способ дообучения современных моделей перевода без использования дополнительных языковых моделей;
2. В работе предложен новый метод переупорядочивания переводческих гипотез, в котором впервые ранжирования переводов, выполненные человеком, используются для борьбы с систематическими ошибками модели. Данный способ дает возможность более дешевого улучшения модели в задаче перевода специализированных текстов, для которых сбор параллельных данных затруднен;
3. В работе впервые предложена интеграция метода маскировки входа в процесс обучения моделей перевода и показано, что такое изменение процесса обучения усложняет процесс обучения, что позволяет улучшить качества в решаемой задаче по сравнению с обучением с авторегрессионной функцией потерь. Показано, что данный метод позволяет адаптировать модели перевода к использованию одноязычных обучающих данных и естественным образом масштабируется для построения моделей постредактирования;
4. В данной работе впервые предлагается метод улучшения тематических моделей за счет использования сегментной структуры документа. Показано, что предложенный соискателем метод решения задачи позволяет улучшить качество контекстного перевода по сравнению с другими методами сегментации.

Практическая значимость

Практическая значимость 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.

В первую очередь, предложенный в работе метод совместного обучения с использованием обратной модели открывает возможность для значительного повышения точности перевода в условиях нехватки параллельных текстов. Такой подход может быть напрямую внедрён в существующие промышленные системы машинного перевода, что позволяет расширять их поддержку на новые языки и специализированные направления без существенных затрат на создание новых корпусов. Это особенно актуально для развития переводческих технологий для языков народов России, национальных меньшинств, а также для задач локализации цифровых продуктов на международные рынки.

Вторая важная составляющая практической значимости связана с преодолением типовых систематических ошибок машинного перевода, таких как недостаточные или избыточные переводы отдельных фрагментов исходного текста. В работе предлагается использовать метод переупорядочивания гипотез на основе разметки, выполненной человеком, — эта технология даёт возможность оперативно дообучать переводческие модели под требования конкретной предметной области, даже когда эталонных параллельных текстов для неё нет. Процесс адаптации систем становится дешевле, количественно сокращаются ошибки, требующие ручной корректировки, а качество выходного перевода приближается к результатам профессиональных переводчиков. Такой подход востребован в электронной коммерции, в автоматизации технического, медицинского и юридического перевода, где важна точность передаваемых сведений и экономия трудозатрат на редактуру текстов. Наиболее важным свойством предложенного подхода является адаптивность. Предложенный способ улучшения качества перевода позволяет автоматически встраивать любые человеческие предпочтения в систему перевода, что существенно сокращает путь до решения задачи автоматического перевода в изначальной постановке.

Особое внимание в работе уделено задаче генерации более лингвистически и грамматически корректных переводов, что достигается с помощью новых функций потерь, в частности, метода маскировки входа. Такой подход не только уменьшает количество грубых ошибок в генерациях, но и позволяет использовать для обучения модели одноязычные корпуса, которые доступны в гораздо больших объёмах по сравнению с параллельными текстами. Это открывает перспективу дальнейшего удешевления внедрения переводческих технологий в бизнес-процессы и государственные сервисы многоязычной поддержки. Предложенные в работе идеи упрощают создание автоматических моделей постредактирования, которые являются неотъемлемой частью CAT-инструментов у профессиональных переводчиков.

Важный практический эффект связан с задачей работы с длинными и сложно структурированными текстами (документами, отчетами, книгами, вебсайтами). Предложенная технология тематической сегментации, основанная на

тематическом моделировании, позволяет разбивать такие тексты на логически цельные фрагменты, что существенно улучшает качество контекстного перевода, облегчает задачу локализации и поиска релевантных фрагментов информации. Эта технология востребована в корпоративном документообороте, в судебной лингвистике, научно-техническом переводе и образовательных порталах — везде, где важно обеспечить согласованный перевод материала на уровне целых блоков информации, а не только отдельных предложений.

Таким образом, результаты диссертационной работы обладают высокой прикладной ценностью как для разработчиков систем автоматического перевода, так и для корпораций, ориентирующихся на автоматизацию многоязычных коммуникаций, локализации цифровых продуктов, повышения доступности информации для широкой аудитории и снижения издержек на ручную работу с текстами. Применение предложенных в работе методов способствует цифровой трансформации различных отраслей, развитию глобального научного и образовательного обмена, поддержке национальных языков и формированию единого информационного пространства.

Достоверность полученных результатов обеспечивается теоретическими обоснованиями, приводимыми соискателем при построении моделей, а также результатами экспериментов, проведённых соискателем.

Основные положения, выносимые на защиту:

1. **Разработан новый вероятностный метод совместного обучения прямой и обратной моделей перевода.** В отличие от предыдущих работ, для данного метода представлен теоретический вывод метода обучения из задачи максимизации правдоподобия циклических переводов. Кроме этого, предложен способ адаптации подхода для обучения современных тяжелых моделей перевода без использования вспомогательных языковых моделей за счет перехода к обучению предобученных моделей. Показано, что включение обратной модели в функцию потерь обеспечивает улучшение качества перевода по сравнению с базовыми методами с точки зрения метрик BLEU и CycleBLEU, особенно в условиях дефицита параллельных корпусов;
2. **Для борьбы с систематическими ошибками перевода предложен и реализован метод переупорядочивания гипотез перевода на основе человеческой разметки.** Данный подход в отличие от предыдущих работ впервые объединяет методологию контрастного обучения с использованием человеческой разметки. Экспериментально показано, что такой способ обучения существенно сокращает долю систематических ошибок (недостаточных, избыточных, некорректных переводов) и заметно эффективнее для быстрой доменной адаптации при отсутствии эталонных переводов;
3. **Для улучшения качества перевода разработан метод маскировки входа для обучения моделей.** В данном подходе впервые идеи маскировки входа применены к задаче перевода и показано, что модификация

вероятностной модели и функции потерь приводит к росту метрик качества автоматического перевода, снижению числа лингвистических и семантических ошибок. Также данный подход позволяет интегрировать обучение на одноязычных данных и на задачу постредактирования в единую модель;

4. **Для улучшения качества документного перевода внедрен и экспериментально обоснован способ тематической сегментации длинных текстов на основе аддитивных тематических моделей.** Для этого при построении тематических моделей впервые использована сегментная структура документа для постобработки E-шага и вывода регуляризатора. Показано, что построенная таким образом тематическая модель лучше подходит для задачи сегментации текстов. Показано, что семантически мотивированная сегментация улучшает качество контекстного перевода длинных документов по сравнению со случайной сегментацией с точки зрения метрики BLEU.

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», Сборник тезисов, г. Москва, 2022
4. Международная конференция «Интеллектуализация обработки информации 2024», тезисы докладов 15-й международной конференции, 2024.

Личный вклад автора. В работах [1], [2] вклад соискателя был определяющим. В публикации [3] соискатель является единственным автором. В работе [4] — разработка и реализация аддитивного сегментирующего регуляризатора в библиотеке BigARTM, получение формулы постобработки E-шага из требований разреживания сегментов, обучение тематических моделей и постановка экспериментов на корпусе PostScience. В работе [5] — обучение тематических моделей ARTM с использованием сегментирующего регуляризатора для последующих экспериментов в задаче поиска. В работе [6] — проведение экспериментов с адаптацией метода переупорядочивания гипотез [1] на основе разметки корпуса предложений, выполненной человеком, на языковой модели.

Соответствие специальности. Тема и содержание диссертации соответствуют специальности 2.3.8 и, в частности, пунктам 4, 5 и 16 паспорта этой специальности.

Публикации. Основные результаты по теме диссертации изложены в 6 печатных изданиях, 2 из которых изданы в журналах, рекомендованных ВАК, 6 — в периодических научных изданиях, индексируемых Scopus.

Объем и структура работы. Диссертация состоит из введения, пяти глав, заключения и списка литературы. Полный объем диссертации **123** страницы

текста с **8** рисунками и 21 таблицами. Список литературы содержит **59** наименований.

Во введении сформулированы основные цели и задачи, описаны основные результаты и структура диссертационной работы.

В первой главе описаны основные вехи развития машинного перевода, а также приводится описание основных механизмов и методов, используемых в обучении современных моделей машинного перевода.

Глава посвящена эволюции генеративных моделей в машинном переводе. Вначале рассматривается история становления этой области: первые проекты автоматизации перевода, появившиеся ещё в 1930-х годах, опередили своё время и стали прообразами современных систем обработки текста.

В дальнейшем развитие машинного перевода было тесно связано с внедрением статистических методов и появлением крупных параллельных корпусов текстов, которые позволили создавать более универсальные и быстрые переводческие системы. Классическим примером здесь стал инструмент Moses, использующий вероятностное моделирование на основе анализа больших коллекций двуязычных текстов и методы эффективного поиска оптимального перевода. Однако такие системы оставались зависимыми от доступности большого количества данных.

Существенный прорыв произошёл с внедрением нейронных моделей, применяемых к задаче перевода с конца 1980-х годов, а в 2010-х — с появлением комплексных нейросетевых архитектур, способных моделировать языковые последовательности без жёсткой формализации правил и грамматик. Особенно значимой вехой стало появление механизма «внимания», который позволил существенно повысить качество перевода длинных и сложных предложений, а также появление архитектуры Transformer.

Далее, в работе формализуется современная постановка задачи машинного перевода как задача обучения параметризованной вероятностной модели, максимизирующей правдоподобие переводного текста при условии исходного на основе параллельных корпусов. В современных системах, как правило, используются авторегрессионные подходы, в которых перевод генерируется итеративно. Архитектура моделей в данной диссертационной работе строится на базе Transformer, общепринятом в области анализа и генерации текстов. Данная архитектура объединяет многослойные кодировщик и декодировщик, содержащие слои внимания внутри. Обучаются данные модели с помощью оптимизатора Adam.

В работе уделено внимание оценке качества в задаче машинного перевода. Для этого используется автоматическая метрика BLEU, учитывающая n-граммные совпадения между переводом модели и эталонными переводами. Эта метрика обладает высокой корреляцией с оценками профессиональных переводчиков и принята как основной стандарт в многочисленных исследованиях, включая эксперименты, представленные в данной работе.

Во второй главе рассматривается подход к решению проблемы нехватки данных при обучении моделей перевода благодаря совместному обучению с обратной моделью перевода.

Одним из методов решения проблемы нехватки параллельных данных является использование синтетических обратных переводов. Обратная модель позволяет использовать синтетические переведенные данные при обучении, получаемые переводом текстов целевого языка на язык входа. Одноязычных документов существенно больше чем параллельных, и с их помощью можно улучшить качество перевода на сложных доменах, в которых тяжело найти хорошие переводы для обучения. Однако такое использование обратной модели не решает проблему ошибок в параллельных обучающих данных.

В данной работе описан способ дообучения модели перевода с использованием информации из обратной модели. Для такого обучения предложен теоретический вывод модели из задачи максимизации правдоподобия циклических переводов:

$$\log P_{\text{cycle}}(x|x) = \log \mathbb{E}_{y \sim P_\Theta(y|x)} P'(x|y) \longrightarrow \max_\Theta,$$

где P — прямая модель, P' — обратная. При этом, предложенный соискателем подход применяется к уже обученной модели перевода. Это позволяет существенно повысить стабильность процесса обучения в сравнении с предыдущими подходами к решению этой задачи, а также позволяет избавиться от использования вспомогательных моделей для стабилизации обучения. Кроме того, использование уже обученной модели при использовании метода существенно сокращает время обучения модели перевода, что делает его применение на практике более эффективным.

Для решения поставленной задачи методом спуска в работе выводится градиент функции потерь циклического перевода с помощью метода REINFORCE:

$$\nabla_\Theta L(x) \approx \log P'(x|y) \nabla_\Theta \log P_\Theta(y|x), \quad y \sim P_\Theta(y|x).$$

Эксперименты как на англо-финском направлении, так и на русско-казахском показали прирост метрики BLEU при использовании метода совместного обучения, несмотря на использование синтетических данных в процессе обучения.

Модель	en-fi, BLEU	ru-kk, BLEU
Без дообучения	23.4	19.5
Прямая	25.0	19.80
Прямая+обратная	**25.5**	**20.05**

В исследовании соискателя также оценивается влияние использования обратной модели перевода на согласованность переводов прямой и обратной моделей с помощью метрики CycleBLEU. В экспериментах удается показать, что

использование обратной модели перевода в обучении позволяет вырастить метрику CycleBLEU. Однако ее прирост становится значительно меньше, если в процессе обучения моделей перевода использовались синтетические обратные переводы.

В третьей главе рассматривается подход по борьбе с систематическими ошибками перевода, такими как избыточные, недостаточные и неточные переводы. Для этого соискателем предлагается метод переупорядочивания гипотез с использованием разметки, выполненной человеком.

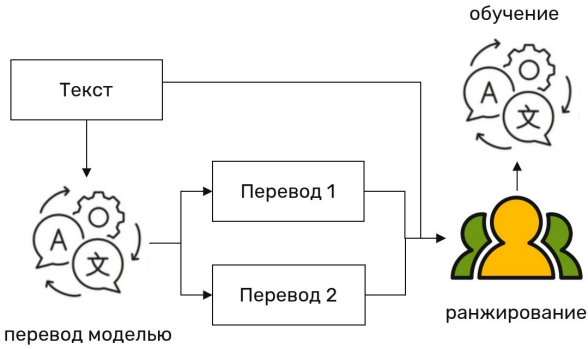
Обучение алгоритмов, лежащих в основе систем перевода, требует большого количества параллельных текстов на разных языках и существенно зависит от качества этих данных и степени их выравненности. Из-за высокой стоимости работы профессиональных переводчиков данные для обучения алгоритмов перевода собираются автоматически с помощью эвристических алгоритмов. При этом высокая степень выравненности отдельных обучающих примеров не гарантируется, что позволяет собирать большие объёмы параллельных текстов.

Алгоритмы перевода при обучении воспроизводят поданные на вход обучающие примеры. Плохо выравненные примеры приводят к появлению систематических ошибок перевода: недопереводов и избыточных переводов. При таких ошибках в переведённом тексте либо теряется часть исходного смысла, либо добавляется какой-то новый. Для борьбы с этими ошибками можно использовать контрастное обучение с негативными примерами. Улучшение качества перевода в таком подходе достигается за счёт увеличения вероятности некоторого перевода y_+ предложения x относительно более плохого перевода y_- того же предложения x с помощью максимальной целевой функции интервала:

$$L(x, y, y_-) = \max(0, \log P_\theta(y_-|x) - \log P_\theta(y|x) + \alpha).$$

Обычно для контастного обучения предлагается получить более плохой перевод y_- из перевода y_+ с помощью эвристических аугментаций. Однако подход с синтетическими примерами позволяет бороться только с определёнными видами ошибок перевода, заданными с помощью эвристики в аугментации. В то же время сложность генерируемых примеров также ограничена сложностью эвристики. Всё это ограничивает широкое применение данного подхода и снижает эффект от улучшения ранжирования. В данной работе соискателем предлагается решение с использованием разметки переводов по качеству, выполненной людьми. В таком подходе y_+ и y_- для обучения модели получаются методом сравнения пары переводов модели человеком. Сами переводы для разметки людьми берутся с помощью разнообразного поиска в ширину при генерации из самой модели перевода. При таком подходе люди оценивают те ошибки перевода, которые модель сама с большой вероятностью готова совершить, что повышает эффективность процесса. Детали процесса получения обучающих данных можно увидеть на картинке ниже.

Экспериментально подтверждается, что данный подход позволяет заметно улучшить качество перевода в среднем с точки зрения метрики BLEU, а



также позволяет значимо уменьшить количество недопереводов и избыточных переводов с точки зрения разметки, выполненной человеком. При этом данный результат достигается не только за счет обучения на победителя разметки, но и за счет использования информации из негативных примеров.

Модель	ru-en-wmt19, BLEU
Базовая модель перевода	35.6
Дообучение на победителя разметки	36.7
Контрастное дообучение на разметку	**37.3**

Кроме того, благодаря предложенной процедуре улучшения с разметкой, выполненной человеком, удалось добиться более высокого качества доменной адаптации модели под языковой домен, на котором не нашлось большого количества параллельных данных — домен электронной коммерции. Предложенный соискателем подход позволил избежать дорогостоящей работы по составлению эталонных переводов для улучшения качества модели перевода на этом домене.

В современных моделях перевода активно используются наработки из области больших языковых моделей. В данной главе также рассматривается использование метода переупорядочивания гипотез на основе разметки, выполненной человеком, в применении к большим языковым моделям. Показано, что прирост в качестве переносится на большие языковые модели и естественным образом масштабируется на задачу контекстного перевода.

Модель	BLEURT	COMET
Языковая модель без дообуч.	73.2	83.1
Языковая модель с дообуч. на разметку	**75.4**	**85.1**

Обученная в том числе с использованием метода переупорядочивания гипотез на основе человеческой разметки языковая модель перевода была представлена на соревновании WMT24. Система «Yandex» на направлении перевода с английского на русский находится в диапазоне с 3 по 7 место, проигрывая по

качеству профессиональному переводчику, а также системе Claude 3.5. При этом модель оказалась лучше 7 систем от других участников соревнования.

В четвёртой главе предлагается метод улучшения качества перевода и борьбы с систематическими ошибками с помощью изменения функции потерь при обучении.

В моделях BERT, mT5, BART и других отказались от решения задачи языкового моделирования в пользу маскированного языкового моделирования. Данный подход показал улучшения качества на многих прикладных задачах анализа текстов. Это может говорить о том, что внутренние представления, выучиваемые в ходе обучения с маскированными функциями потерь лучше представляют семантику и контекстные связи внутри текста. В предложенном соискателем подходе предлагается применить метод маскировки на задачу машинного перевода.

Для этого маскированная функция потерь перевода записывается следующим образом:

$$L_\theta(D) = \sum_{i=1}^{M} \sum_{j=1}^{|y_i|} \log P_\theta(y_i^j | y_i^{<j}, m(y_i), m(x_i)).$$

При обучении методом маскировки модель при генерации перевода видит входной и целевой тексты, в которых значительная часть последовательных токенов замаскирована. Маскирование входного текста заставляет модель внимательнее смотреть на те слова входа, которые остались доступны. Кроме этого, добавление маскированного перевода во вход модели позволяет учитывать правый контекст, что должно ограничивать зависимость генерации каждого нового слова от левого контекста.

В результатах экспериментов отмечено, что обучение модели перевода с маскированной функцией потерь дает улучшение по метрике BLEU по сравнению с авторегрессионным бейзлайном. Кроме того, показано, что предложенный соискателем подход позволяет расширить решаемую задачу и включить одноязычные данные в процесс обучения модели перевода, однако в экспериментах добавление одноязычных данных на русском языке не принесло качественного результата в задаче перевода. Это может объясняться тем, что англо-русское направление перевода не является малоресурсным. В будущих исследованиях предполагается, что добавление одноязычных данных может лучше себя проявить на направлениях, где параллельных данных для обучения меньше.

Модель	en-ru-wmt19, BLEU
Авторегрессионная функция потерь	34
Функция потерь маскирования	**34.35**
Функция потерь маск. с одноязыч. данными	33.5

Кроме того, предложенный соискателем подход позволяет адаптировать модель к решению задачи постредактирования, где по исходному тексту и первичному переводу нужно авторегрессионно построить его улучшенную версию. В экспериментах удается показать, что качество постредактирования переводов слабой модели заметно растет при использовании описанного подхода.

Модель	en-ru-wmt19, BLEU
Слабая модель перевода (СМП)	30.3
Модель постредактирования СМП	**34.1**

В пятой главе рассматривается влияние качества тематической сегментации на итоговое качество машинного перевода длинных документов. Несмотря на значительный прогресс в современных моделях машинного перевода, задача перевода длинных, тематически неоднородных текстов остаётся сложной. Разбиение таких текстов на произвольные или недостаточно однородные сегменты разрушает контекстные связи и негативно влияет на качество перевода. Для решения этой проблемы предлагается использовать тематические модели ARTM, позволяющие автоматически делить текст на смысловые фрагменты, внутри которых сохраняется единство тематики. Такой подход обеспечивает более согласованный перевод и улучшает связность итогового текста.

Излагается теоретическая база тематических моделей, приводится описание аддитивной регуляризации и предлагается улучшение ЕМ-алгоритма, позволяющее учитывать внутридокументную сегментную структуру. Для интеграции сегментной структуры документов и адаптации к ней тематических моделей предлагается процедура постобработки Е-шага:

$$\widetilde{p_{tdw}} = p_{tdw}\left(1 - \frac{\tau}{n_{dw}}\sum_{s\in S_d}\frac{n_{sw}}{n_s}\left(\frac{1}{p_{ts}} - \sum_{z\in T}\frac{p_{zdw}}{p_{zs}}\right)\right),$$

где S_d — множество всех предложений в документе.

Предложенный соискателем метод постобработки Е-шага выводится из предположения, что внутри одного сегмента текста распределение тематик должно быть как можно дальше от равномерного. При этом, при обучении тематической модели границы сегментов пересчитываются на каждой итерации с помощью эвристического алгоритма из TopicTiling. Данный метод работает на основе сравнения близости тематических векторов слева и справа от предполагаемой границы сегмента.

Обучение такой модели дает улучшение с точки зрения метрик сегментации. Из этого делается вывод о полезности использования сегментной структуры документа при построении тематических моделей.

Кроме этого исследуется практическая полезность тематической сегментации в задаче информационного поиска: сегментация по темам облегчает сопоставление сегментов документов из поиска и запроса. Эксперименты показывают, что тематическая сегментация обеспечивает хорошие результаты

Модель	**WindowDiff**	P_k
TopicTiling	0.258	0.145
PLSA + разреженность Θ	0.173	0.100
PLSA + SentenceSparse	0.159	0.099
PLSA + SegmentSparse	0.155	0.095

поиска с точки зрения точности в сравнении с другими методами построения текстовых представлений. Это является дополнительным подтверждением, что семантическая сегментация на основе тематических моделей строит полезные в практических задачах текстовые сегменты.

В финальной части главы результаты тематической сегментации применяются непосредственно к задаче контекстного машинного перевода длинных текстов. Тематически однородные сегменты переводятся отдельными блоками, что позволяет повысить качество перевода в сравнении со случайной сегментацией. Проведённые на тестовых данных WMT22 эксперименты демонстрируют, что предварительная тематическая сегментация приводит к приросту метрики BLEU.

Модель сегментации	WMT22, BLEU
Случайная сегментация	25.1
Сегментация PLSA+SegmentSparse	**25.4**

Тематические модели на базе ARTM оказываются эффективными и относительно простыми для внедрения, они позволяют автоматически регулировать детализацию сегментации, что может быть востребовано в реальных рабочих процессах переводчиков. Эти результаты подтверждают высокую значимость качества сегментации для перевода длинных документов и демонстрируют преимущества интеграции современных методов тематического моделирования в технологический стек перевода.

В заключении формулируются выводы, полученные в данном диссертационном исследовании, а также задаются направления дальнейших исследований.

В рамках данного исследования соискателем рассмотрены различные ограничения систем машинного перевода, которые на практике приводят к низкому качеству автоматического перевода.

Основные результаты работы заключаются в следующем:

1. Разработан новый вероятностный метод совместного обучения прямой и обратной моделей перевода. Обучены модели перевода на англо-финском и русско-казахском языковых направлениях с применением предложенного метода. Показано, что использование обратной модели в обучении дает улучшение с точки зрения метрик качества машинного перевода: BLEU и CycleBLEU. Кроме этого, предложен способ адаптации подхода для обучения современных тяжелых моделей перевода без использования вспомогательных языковых моделей;

2. Предложен и реализован метод переупорядочивания гипотез перевода на основе человеческой разметки. Обучена модель перевода с русского на английский с помощью предложенного метода. Экспериментально показано, что такой способ обучения существенно сокращает долю систематических ошибок, таких как недостаточные, избыточные и некорректные переводы. Показано, что предложенный соискателем метод эффективен для быстрой доменной адаптации при отсутствии эталонных переводов. Показано, что данный метод успешно переносится на дообучение современных языковых моделей, адаптированных под задачу перевода;
3. Разработан метод маскировки входа для обучения моделей. Обучена модель перевода с английского на русский с помощью предложенного метода. Экспериментально показано, что модификация вероятностной модели и функции потерь приводит к росту метрики BLEU, а также снижению числа лингвистических и семантических ошибок. Также данный подход позволяет интегрировать обучение на одноязычных данных, на задачу перевода и на задачу постредактирования в единую модель;
4. Разработан и экспериментально обоснован способ тематической сегментации длинных текстов на основе аддитивных тематических моделей. Для этого при построении тематических моделей использована сегментная структура документа и реализована постобработка E-шага. Экспериментально показано, что построенная таким образом тематическая модель применима в задачах сегментации текстов и информационного поиска. Показано, что семантически мотивированная сегментация улучшает качество контекстного перевода длинных документов по сравнению со случайной сегментацией с точки зрения метрики BLEU.

Перспективные направления дальнейшей работы связаны, прежде всего, с расширением возможностей предложенных методов и их адаптацией к новым технологическим и прикладным вызовам. Одной из ключевых задач является интеграция современных больших языковых моделей (LLM) в задачу машинного перевода с применением разработанных в диссертации подходов — таких как совместное обучение с обратной моделью и обучение с функциями потерь, ориентированных на человеческие предпочтения. Это позволит не только повысить качество перевода за счёт обобщающей мощности LLM, но и сделать переводческие системы более управляемыми и устойчивыми к ошибкам.

Важным направлением остается дальнейшее развитие методов адаптации для малоресурсных языков, а также экспериментальное исследование эффективности предложенных решений при обучении на смешанных и мультиязычных корпусах, включая экстремальные случаи, когда объем параллельных данных минимален. В данном направлении предлагается адаптация предложенных подходов к обучению мультиязычных систем, где разные родственные языки могут

компенсировать нехватку данных на определенных направлениях. Перспективно выглядит применение тематической сегментации и вероятностных методов не только для улучшения перевода длинных документов, но и для задач суммаризации, автоматической разметки и автоматического редактирования текстов.

Наконец, существенный интерес представляет создание комплексных систем автоматического перевода, сочетающих перечисленные в работе методы в едином пайплайне: с возможностью интерактивной обратной связи от человека, доменной адаптации «на лету» и постоянного самообучения на реальных пользовательских и корпоративных данных. Разработка таких систем будет способствовать дальнейшему повышению качества, гибкости и доверия к автоматическому машинному переводу в различных сферах применения.

Публикации автора по теме диссертации

1. *Воронцов*, *К.* Упорядочивание гипотез в моделях перевода с использованием человеческой разметки [Текст] / К. Воронцов, Н. Скачков // Известия Российской Академии Наук. Теория и системы управления. — 2024. — № 4. — С. 121—128. — Vorontsov K.V., Skachkov N.A. Reranking Hypotheses in Translation Models Using Human Markup // Journal of Computer and Systems Sciences International. 2024. V. 63. No 4. PP. 679-686. — (ВАК, Scopus).
2. *Скачков*, *Н. А.* Улучшение качества машинного перевода с использованием обратной модели [Текст] / Н. А. Скачков, К. В. Воронцов // Автоматика и телемеханика. — 2022. — № 12. — С. 31—43. — N.A. Skachkov, K.V. Vorontsov. Improving the Quality of Machine Translation Using the Reverse Model // Automation and Remote Control. 2022. Vol. 83. PP. 1897-1907. — (ВАК, Scopus).
3. *Skachkov*, *N.* Method of Input Masking for Training Translation Models [Text] / N. Skachkov // Pattern Recognition and Image Analysis. — 2025. — Vol. 35. — P. 493—500. — (Scopus).
4. *Skachkov*, *N. A.* Improving Topic Models with Segmental Structure of Texts [Text] / N. A. Skachkov, K. V. Vorontsov // Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference ”Dialogue 2018”. — Moscow, 2018. — Vol. 17. — P. 652—661. — (Scopus).
5. Applying Topic Segmentation to Document-Level Information Retrieval [Text] / G. Shtekh [et al.] // Proceedings of the 14th Central and Eastern European Software Engineering Conference Russia. — Moscow, Russian Federation, 2018. — (CEE-SECR ’18). — (Scopus).
6. From General LLM to Translation: How We Dramatically Improve Translation Quality Using Human Evaluation Data for LLM Finetuning [Text] / D. Elshin [et al.] // Proceedings of the Ninth Conference on Machine Translation. — Miami, Florida, USA, 2024. — Nov. — P. 247—252. — (Scopus).