Федеральный исследовательский центр “Информатика и управление”
Российской академии наук

На правах рукописи

Скачков Николай Андреевич

Вероятностные нейросетевые модели понимания и генерации текстов естественного языка

Специальность 2.3.8 —
«Информатика и информационные процессы»

Диссертация на соискание учёной степени
кандидата технических наук

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

Москва — 2026

Оглавление

Стр.

Стр.

Стр.

Введение

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 машинном переводе, 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 избыточ-

ных, 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 развитии архитектур последовательных моделей с механизмом внимания и трансформеров был достигнут в работах [1], [2]. Далее архитектурные изменения моделей и методов перевода шли по пути решения конкрентных прикладных проблем. Так, в работе [3] авторами был предложен подход, основанный на стохастической обработке входной последовательности, позволяющий лучше выучивать представления редких входных подслов. В работах [4], [5] авторы предлагают неавторегрессионный подход к генерации переводов, что позволяет 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 вклад в решение этой задачи с помощью интеграции обратной модели перевода в обучение был внесен в работе [6]. 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. В указанной работе авторы предложили компенсировать нехватку параллельных данных с помощью генерации синтетических данных обратной моделью. Идеи совместного обучения прямой и обратной моделей перевода были предложены в работе [7]. Тем не менее, предложенные ими алгоритмы требуют хранения в памяти дополнительных языковых моделей, что делает их применение для современных масштабных архитектур вычислительно неэффективным.

Проблема борьбы с систематическими ошибками нейронных моделей, такими как пропуски слов и избыточная генерация, исследовалась в работе [8]. Авторами было предложено использовать контрастное обучение с негативными примерами. Однако в существующих подходах негативные примеры формируются искусственно на основе жестких эвристических шаблонов, что не позволяет

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, представленный в моделях BERT [9], BART [10], mT5 [11] и других. За счет изменения процесса обучения и усложнения функции потерь, используемой при обучении, авторам 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 исследований.

Проблемы контекстного перевода были подняты в работе [12]. 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. Для решения задачи семантической сегментации могут использоваться тематические модели [13], [14]. Например, в работах [15] и [16] тематические модели используются для сегментации текстов благодаря учету семантически однородной структуры текста при построении тематических моделей. 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 [17].

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.

Целью 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 [18], BLEURT [19], COMET [20], а также разметка переводов на качество, выпоненная 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, веб-сайтами). 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 зищиту:

1. **Разработан новый вероятностный метод совместного обучения прямой и обратной моделей перевода.** В отличии 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 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 [18] 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 отличии 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 пост-обработки 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 [18].

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 [54], [55] вклад соискателя был определяющим. В публикации [56] соискатель является единственным автором. В работе [57] — разработка и реализация аддитивного сегментирующего регуляризатора в библиотеке BigARTM [21], получение формулы постобработки E-шага из требований разреживания сегментов, обучение тематических моделей и постановка экспериментов на корпусе PostScience. В работе [58] — обучение тематических моделей ARTM [13] с использованием сегментирующего регуляризатора для последующих экспериментов в задаче поиска. В работе [59] — проведение экспериментов с адаптацией метода переупорядочивания гипотез [54] 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.

Работа состоит из введения, пяти глав, заключения и списка литературы.

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. 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.

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

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. На русско-казахском также наблюдается рост метрики согласованности переводов прямой и обратной моделей.

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 обучение с негативными примерами [8]. 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.

Авторы статьи предлагают получить более плохой перевод y_- из перевода y_+ с помощью эвристических аугментаций [8]. 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.

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.

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.

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

Для этого рассматривается метод маскировки, который дает улучшение качества на многих прикладных задачах анализа текстов по сравнению с классическими методами обучения [11].

В работе предлагается адаптировать метод маскирования входа для задачи перевода.

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 другими подходами. Предложенный 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.

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.

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 [13], 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 EM-алгоритма, позволяющее учитывать внутридокументную сегментную структуры. Для интеграции сегментной структуры документов и адаптации к ней тематических моделей предлагается процедура постобработки E-шага.

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 поиска с точки зрения точности.

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. 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.

Объем и структура работы. Диссертация состоит из введения, 5 глав и заключения. Полный объём диссертации составляет 123 страницы, включая 8 рисунков и 21 таблицу. Список литературы содержит 59 наименований.

Глава 1. Основные понятия и обзор развития генеративных моделей машинного перевода

1.1 История развития области

1.1.1 Ранние подходы к постановке задачи машинного перевода

Машинный перевод (МП) представляет собой одно из наиболее значимых достижений современной компьютерной лингвистики, открывшее новую эру в развитии межъязыковой коммуникации. История его возникновения и развития неразрывно связана с эволюцией вычислительной техники и стремлением человечества преодолеть языковой барьер в эпоху глобализации. Появление первых идей о возможности автоматизации процесса перевода относится к периоду зарождения кибернетики, когда учёные начали активно исследовать возможности моделирования человеческого мышления с помощью вычислительных машин.

Предпосылки создания систем машинного перевода формировались на стыке нескольких научных дисциплин: лингвистики, математики, кибернетики и информатики. Стремительное развитие вычислительной техники в середине XX века создало необходимые технические условия для реализации амбициозной задачи автоматического перевода текстов. Первые эксперименты в этой области были предприняты вскоре после появления электронно-вычислительных машин, что положило начало новому направлению в науке о языке и открыло широкие перспективы для международного сотрудничества и обмена информацией между представителями различных языковых сообществ.

В 1933 году П. П. Смирнов-Троянский представил в Академию наук СССР собственную разработку — устройство, предназначенное для автоматизации подбора и печати слов при переводе текста с одного языка на другой. Конструкция представляла собой стол с наклонной поверхностью, над которой был закреплён фотоаппарат, работающий синхронно с пишущей машинкой. На стол был помещён так называемый «глоссарный слой» — передвижная пластина, на которой были напечатаны слова на трёх, четырёх или большем количестве языков. На это изобретение Смирнов-Троянский получил авторское свидетельство. Однако идея

Смирнова-Троянского была встречена академическим сообществом с недоверием и вскоре оказалась забыта на долгое время. Лишь в 1959 году интерес к его устройству возродился после публикации И. К. Бельской и Д. Ю. Пановым сборника материалов, посвящённого машине для автоматического перевода, предложенной Смирновым-Троянским в 1933 году.

Одновременно с Петром Смирновым-Троянским «машину для машинного перевода» пытался запатентовать французский изобретатель Жорж Артцруни, однако идеи первого были более проработанными — система включала в себя, кроме двуязычного словаря, схему для кодирования межъязыковых грамматик на основе эсперанто и общие идеи анализа и синтеза текста.

Первой успешной демонстрацией возможностей математических моделей в переводе был «Джорджтаунский эксперимент». Он прошел 7 января 1954 года в Нью-Йорке в центральном офисе IBM. Организаторами выступили Джорджтаунский университет при сотрудничестве с IBM. В рамках эксперимента была продемонстрирована полностью автоматизированная система, осуществившая перевод более 60 фраз с русского языка на английский. Это событие оказало значительное влияние на последующее развитие исследований в области машинного перевода на протяжении последующих двенадцати лет.

Цель эксперимента заключалась в том, чтобы привлечь внимание широкой общественности и государства к потенциалу машинного перевода. Интересно, что для демонстрации использовалась достаточно простая система: её языковая модель включала всего 6 грамматических правил, а словарь насчитывал около 250 лексических единиц. Программа была узкоспециализированной: основной предметной областью стала органическая химия, однако были и предложения общего содержания. Для работы использовался мейнфрейм IBM 701, куда предлагалось вводить, например, такие предложения, как: «Обработка повышает качество нефти» или «Командир получает сведения по телеграфу». Система воспринимала их с перфокарт и выдавала перевод на английском языке, распечатанный транслитом.

Руководителями эксперимента были профессор Леон Достер из Джорджтаунского университета и глава отдела прикладных исследований IBM Катберт Хёрд. За лингвистическую часть отвечал адъюнкт-профессор Пол Гарвин, а компьютерное обеспечение обеспечивал Питер Шеридан из IBM.

Демонстрация получила широкое освещение в прессе и была воспринята общественностью как значительный прогресс, что способствовало увеличению государственных инвестиций в вычислительную лингвистику, особенно в США.

В том же 1954 году аналогичный эксперимент по созданию машинного перевода был проведён и в Советском Союзе, в Институте точной механики и вычислительной техники АН СССР на компьютере БЭСМ. Руководила исследованиями Изабелла Бельская, а инициировал проект директор института Дмитрий Панов. Параллельно над подобными задачами работала группа учёных в Математическом институте имени В. А. Стеклова АН СССР под руководством Ольги Кулагиной и Алексея Ляпунова.

Организаторы эксперимента в Джорджтауне заявляли, что проблемы автоматического перевода удастся решить за 3–5 лет. Однако эти прогнозы не оправдались: к 1966 году комиссия ALPAC пришла к выводу, что более чем десятилетние усилия так и не привели к созданию полноценной системы автоматического перевода, после чего финансирование работ было заметно сокращено.

1.1.2 Статистический машинный перевод

Основания статистического машинного перевода (СМП) заключаются в том, что машинный перевод базируется на статистических моделях, параметры которых выводятся путём анализа больших коллекций двуязычных текстов. Несмотря на то, что первые мысли о статистическом подходе к переводу появились ещё в работах Уоррена Уивера еще в 1949, полноценные статистические системы начали разрабатываться только в начале 1990-х годов, когда IBM инициировала масштабные исследования по этому направлению в своём научном центре имени Томаса Уотсона. Эти многолетние разработки привели к появлению сложных инструментов, которые смогли обеспечить качественный машинный перевод для разных языковых пар.

Ярким примером зрелой системы этого класса можно считать Moses — фреймворк, созданный в Эдинбургском университете в 2017 году. Для каждой фразы на языке оригинала Moses строит набор возможных вариантов перевода, опираясь на условные вероятности появления перевода, вычисленные на основе большого корпуса параллельных текстов. Для быстрого перебора гипотез применяется стратегия поиска в ширину. Кроме того, Moses предусматривает использование дополнительных лингвистических модулей, отвечающих за корректное образование словоформ, их лемматизацию и преобразование в форме,

соответствующей грамматике целевого языка. Подобный подход к переводу отличается высокой скоростью и универсальностью — методы СМП подходят для разных языков и требуют относительно небольших ресурсов по сравнению с системами, основанными на правилах, которые использовались в более ранних подходах. Однако, для их работы необходимо наличие огромных параллельных корпусов, способных обеспечить модели точными статистическими закономерностями. Без них невозможно построение корректных моделей и внутренних представлений для используемых языков.

1.1.3 Нейронные модели понимания и генерации текстов в переводе

Первые опыты применения нейронных сетей в машинном переводе начались еще в конце 1980-х годов. Так, в 1987 году Роберт Б. Аллен продемонстрировал возможности прямых нейронных сетей для перевода автоматически сгенерированных английских предложений — пусть и с очень ограниченным словарём — на испанский язык. Эти системы были весьма примитивными: размер их входного и выходного слоя соответствовал самой длинной фразе в исходном или целевом языке. Вся сложность языка сводилась к небольшому набору комбинаций, однако даже такой опыт знаменателен для истории — в те годы нейронные сети только начинали рассматриваться в роли моделей для смысловой обработки языка.

В дальнейшем, в начале 1990-х, исследователи пытались усложнить архитектуру: появились автоассоциативные сети памяти, способные кодировать последовательности переменной длины в фиксированное представление, а потом обратно восстанавливать фразу по этой «сжатой» информации. Примером может служить работа Лонни Крисмана, который обучал отдельные нейронные сети для исходного и целевого языка, связывая их через общее скрытое представление. Постепенно эти подходы упрощались и совершенствовались, но до настоящего качественного скачка было еще далеко — главным образом из-за нехватки вычислительных мощностей.

Потому в 1990-х и 2000-х годах на первое место вышел статистический машинный перевод, речь про который была в предыдущем разделе. Тем не менее, даже в эпоху «власти статистики» учёные продолжали эксперименты с нейрон-

ными подходами, внедряя их как часть гибридных систем. Нейросети заменяли, например, обычные фразовые языковые модели, или подстраивали оценки для вероятности фраз. В это время шёл поиск оптимального способа соединить статистический, лингвистический и нейросетевой принципы обработки языка.

Ситуация кардинально изменилась в начале 2010-х, когда возникли комплексные нейросетевые системы: нейросетевые переводчики, способные обучаться без жёсткой формализации внутренних правил. Появились архитектуры, моделирующие языковые последовательности: сначала с использованием свёрточных, а потом рекуррентных нейросетей. Но вскоре обнаружились их пределы — такие модели с трудом справлялись с длинными и сложными предложениями.

Значительный прорыв был связан с внедрением механизма внимания. В нем при каждом шаге перевода система могла «фокусироваться» на тех участках исходного текста, которые важны для текущего слова в переводе. Это дало мощный толчок развитию: в 2015–2016 годах такие компании как Baidu, Google выпустили промышленные нейронные переводчики, а на научных конференциях нейротехнологии вывели на новый уровень всю область машинного перевода. Следующий этап — создание трансформеров, предложенных в 2017 году. Эти модели оказались более простыми, они использовали только механизм внимания, но при этом позволяли значительно увеличить размер используемых моделей. К 2020-м годам такие модели стали стандартом в индустрии. DeepL, Microsoft Translator, Yandex.Translate быстро интегрировали эти технологии, а появление моделей уровня GPT-3 показало, что большие языковые модели вообще могут выступать как универсальные переводчики на лету. Практика показала: эффективнее сначала обучать большие языковые модели на гигантских массивах текстов на одном языке, а затем — на параллельных корпусах, что особенно актуально для малоизученных языков. Более того, современные генеративные языковые модели (такие как GPT-3.5 или GPT-4) могут выполнять перевод даже без отдельного дообучения под задачу перевода. Хотя качество при работе с некоторыми языками всё ещё уступает специализированным системам.

1.2 Постановка задачи машинного перевода

С математической точки зрения задача машинного перевода формулируется как задача поиска такого отображения или вероятностного распределения, которое ставит в соответствие каждому входному тексту на входном языке оптимальный или наиболее вероятный перевод на целевом языке. Современные модели ориентированы не только на восстановление точного перевода, но и на оценку степени соответствия возможных вариантов перевода, что позволяет учитывать вариативность естественных языков и неоднозначность перевода отдельных конструкций. Для построения и обучения таких моделей используется корпус параллельных текстов, на основе которого формализуется задача оптимизации соответствующих функций потерь.

Опишем для начала математическую постановку задачи машинного перевода. Классическая задача автоматического перевода формулируется как преобразование текста, принадлежащего одному исходному языку, в эквивалентный по смыслу, грамматике и стилистике текст на другом целевом языке. Пусть дано множество пар текстов на двух языках:

$$\{(x_i, y_i)\}_{i=1}^{N},$$

где x_i — исходные тексты на языке входа (ЯВ), а y_i — соответствующие им переводы на целевой язык (ЦЯ). Задача состоит в том, чтобы построить вероятностную модель перевода $P_\theta(y|x)$, параметризованную вектором параметров θ, которая приближает истинное распределение переводов по корпусу примеров.

В терминах машинного обучения поставленная задача заключается в поиске таких параметров модели θ, при которых вероятность или логарифм вероятности получить правильный перевод y_i по заданному исходному тексту x_i максимальна на обучающей выборке. Формально это записывается в виде задачи максимизации логарифма правдоподобия (MLE):

$$\sum_{i=1}^{N} \log P_\theta(y_i|x_i) \longrightarrow \max_\theta,$$

где суммирование идет по всем примерам обучающего корпуса, а параметры θ подбираются с помощью стохастических методов оптимизации.

Следует подчеркнуть, что в современном машинном переводе требуется не просто оценивать вероятность перевода, но и уметь генерировать близкие по качеству к профессиональным переводы, а также делать это быстро даже при высокой размерности пространства возможных переводов. Для этого используются авторегрессионные или пошаговые вероятностные модели [22], в которых построение перевода происходит итеративно: на каждом шаге порождается следующее слово перевода на основе исходного текста и уже сгенерированного префикса целевого текста.

В таких моделях вероятность генерации полного перевода y по исходному тексту x раскладывается по формуле произведения условных вероятностей по цепному правилу:

$$\log P_\theta(y|x) = \sum_{t=1}^{|y|} \log P_\theta(y^t|y^{<t}, x), \tag{1.1}$$

где y^t — t-ое слово (или токен) перевода, а $y^{<t} = (y^1, \ldots, y^{t-1})$ — уже сгенерированный префикс длины $t-1$. Таким образом, модель «строит» перевод шаг за шагом, дополняя последовательность новым словом, выбранным из словаря целевого языка.

Авторегрессионная постановка имеет ряд теоретических и практических преимуществ. Во-первых, она позволяет избежать полного перебора всех возможных последовательностей слов, чей размер экспоненциально растет с длиной текста. Во-вторых, она естественным образом моделирует языковые и смысловые зависимости между словами в целевом тексте. В-третьих, она позволяет реализовать современные архитектуры нейронных сетей типа кодировщик-декодировщик с механизмом внимания, успешно применяемые для многих задач понимания и генерации текста.

Однако авторегрессионность вводит и определенные ограничения: каждое следующее слово строится по уже сгенерированному префиксу, что требует аккуратного учета накопленных ошибок, эффекта «нарастающего шума», а также усложняет параллельную обработку.

Современные нейросетевые генеративные модели машинного перевода используют именно авторегрессионные принципы, что позволяет им достигать качества, сопоставимого с профессиональным переводом [22].

Важно отметить, что эффективность решения задачи сильно зависит от структуры и объема обучающих данных (корпусов параллельных текстов), слож-

ности языковой пары, выборки жанров и тематик. В частности, для редких языков и специфических доменов (научная, техническая документация, медицина, право) задача машинного перевода приобретает дополнительные теоретические и прикладные сложности.

Таким образом, основная задача обучения моделей нейронного машинного перевода — найти такие параметры модели, которые позволяют максимально точно и полно восстанавливать смысл исходного текста на целевом языке в условиях ограничения на вычислительные ресурсы и доступные данные. В ходе дальнейшей работы под моделью перевода будем подразумевать именно авторегрессионную вероятностную модель 1.1, построенную на современных нейросетевых принципах.

1.3 Архитектура Transformer

В основе современных систем нейронного машинного перевода лежит архитектура Transformer [2], которая совершила прорыв в области обработки последовательностей за счёт нового принципа построения нейронной сети без использования рекуррентных и сверточных элементов. В отличие от предшественников Transformer опирается исключительно на механизм внимания, избегая использования рекуррентных связей, что позволяет эффективно моделировать как краткосрочные, так и долгосрочные зависимости в последовательности без потери связности и с возможностью значительного параллелизма.

1.3.1 Рекуррентные нейронные сети для моделирования текстов

Рекуррентные нейронные сети — это специальный класс нейронных сетей, предназначенный для обработки последовательной информации, где каждый элемент входа зависит не только от текущего входного значения, но и от результатов вычислений на предыдущих шагах [23]. Классическая рекуррентная сеть состоит из повторяющегося блока или ячейки, на каждом временном шаге принимающей на вход текущее значение последовательности и скрытое состояние, сформиро-

ванное на предыдущем шаге. Благодаря этому механизм обмена информацией между шагами, сеть способна аккумулировать и передавать важные признаки последовательности от начала к концу, что делает её естественным выбором для задач перевода, синтеза текста, анализа времени и других задач обработки текстовых и временных данных.

Применение обучаемого рекууретного слоя к последовательности или тексту описывается следующим образом. Пусть x_t — вход на шаге t, h_{t-1} — скрытое состояние на предыдущем шаге, h_t — скрытое состояние на текущем шаге, y_t — выход сети на текущем шаге. Тогда обновление скрытого состояния и вычисление выхода записываются так:

$$h_t = f(W_{xh}x_t + W_{hh}h_{t-1} + b_h),$$
$$y_t = g(W_{hy}h_t + b_y),$$

где

– W_{xh} — матрица весов для входа,
– W_{hh} — матрица весов скрытого состояния,
– W_{hy} — матрица весов для выхода,
– b_h, b_y — смещения (bias),
– $f(\cdot)$ — функция активации скрытого слоя (например, $\tanh$ или ReLU),
– $g(\cdot)$ — функция активации выхода (например, $\mathrm{softmax}$ для задачи классификации).

В процессе обработки последовательности рекуррентная сеть итеративно применяет эти преобразования для каждого элемента последовательности $\{x_1, x_2, \ldots, x_T\}$, последовательно обновляя скрытое состояние h_t:

$$h_t = f(W_{xh}x_t + W_{hh}h_{t-1} + b_h), \quad t = 1, \ldots, T$$

Обычно начальное скрытое состояние h_0 либо инициализируется нулями, либо обучается как отдельный параметр.

Однако в процессе обучения и практического применения у классических RNN выявились существенные ограничения. Во-первых, такое пошаговое перемещение и обновление скрытого состояния приводит к тому, что модель забывает информацию по мере продвижения по последовательности — явление, известное как затухание или взрыв градиентов. Это приводит к тому, что модель слабо усваивает зависимости, лежащие слишком далеко друг от друга по временной оси,

например, связь между началом и концом предложения, что критично для языковых задач.

Для частичного решения этой проблемы были разработаны более сложные разновидности рекуррентных сетей — такие как LSTM и GRU [23]. В них используется специальный механизм ворот для управления потоком информации, позволяющий лучше хранить долгосрочные зависимости в памяти. Однако даже такие сети имеют встроенные ограничения: обработка данных оказывается строго последовательной, и поэтому практически не поддаётся параллелизации. Это серьёзно ограничивает скорость обучения и генерации, особенно на длинных текстах. Визуализацию рекуррентного блока ячейки LSTM можно увидеть на рисунке 1.1.

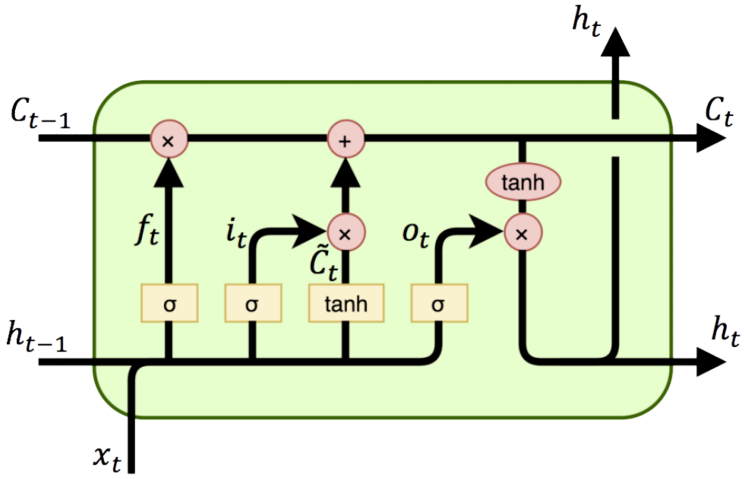


Рисунок 1.1 — Визуализация ячейки LSTM.

Ключевым шагом вперёд в моделировании языковых последовательностей стал механизм внимания, который лег в основу архитектуры Transformer. В отличие от рекуррентных сетей механизм внимания позволяет модели на каждом шаге непосредственно обращаться ко всем остальным позициям входной последовательности, независимо от их удалённости друг от друга. Это даёт возможность мгновенно учитывать контекст всего предложения или документа, эффективно моделируя как близкие, так и дальние зависимости между элементами последовательности. Более того, архитектура Transformer допускает полностью параллельную обработку всех позиций последовательности, что приводит к резкому увеличению производительности, значительному ускорению обучения и большей эффективности работы на больших данных. Благодаря этим аспектам архитектуры с механизмом внимания продемонстрировали значитель-

ный прирост качества в задачах машинного перевода и вытеснили рекуррентные сети из ведущих практик в области обработки естественного языка.

1.3.2 Детали архитектуры Transfrmer

Традиционный Transformer состоит из двух основных частей: кодировщика и декодировщика. Обе части состоят из последовательного применения L одинаковых по структуре блоков-слоёв, каждый из которых содержит несколько ключевых компонент.

Кодировщик принимает на вход последовательность векторных представлений исходного предложения (на входном языке), обрабатывая её через L блоков. Каждый блок включает два основных подслоя. Первый слой — многоуровневый механизм внимания на себя, который позволяет каждому слову учитывать все остальные слова текущего предложения. Второй слой — полносвязная позиционно-независимая нейронная сеть, применяемая к каждому элементу последовательности независимо.

На каждом шаге после основных подслоёв добавляются остаточные связи, результат каждого подслоя нормализуется с помощью дополнительного слоя послойной нормализации.

Декодировщик симметричен кодировщику по архитектуре, однако каждый его блок содержит три подслоя. Первый слой — механизм маскированного внимания на себя, в котором информация доступна только о предыдущих словах в целевом тексте. При этом правый контекст маскируется. Второй слой — механизм внимания на выход кодировщика: декодер на каждом шаге имеет доступ ко всем скрытым состояниям кодировщика. Это позволяет декодеру использовать всю информацию исходного предложения при генерации. Третий слой — аналогичная своя полносвязная сеть и нормализация.

Маскирование в самовнимании реализовано для того, чтобы не было «подглядывания» вперёд при авторегрессионной генерации перевода.

До подачи в Transformer слова преобразуются в эмбеддинги фиксированной размерности. Для сохранения порядка слов в предложении используются позиционные эмбеддинги, которые либо обучаются вместе с моделью, либо задаются априорно, например, синусоидально.

Главным преимуществом Transformer является способность связывать элементы последовательности любой длины напрямую с помощью внимания — каждый выходящий вектор получает информацию от всех входных токенов. Это приводит к значительному снижению проблемы «затухания градиентов» по сравнению с рекуррентными сетями и позволяет масштабировать модели на очень длинные тексты и большие корпуса данных.

Механизм внимания для одной головы внимания формально вычисляется как:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V$$

где Q — матрица запросов, K — матрица ключей, V — матрица значений, d_k — размерность ключа. Механизм многоголового внимания позволяет обучать модель видеть информацию «под разными углами», одновременно реализуя несколько параллельных представлений.

Визуализацию работы архитектуры Transformer можно увидеть на рисунке 1.2.

Обучение нейронных моделей перевода требует эффективного и надёжного метода оптимизации. В качестве такового в большинстве современных экспериментов используется алгоритм Adam [24]. Adam предназначен для стохастической оптимизации функций большого числа переменных и объединяет в себе достоинства методов AdaGrad, то есть сохраняет индивидуальный темп обучения для каждого параметра, и RMSProp, то есть адаптивно изменяет темп обучения на основе экспоненциально взвешенной скользящей средней квадратов градиентов.

Работа Adam основана на поддержании для каждого параметра моментов первого (среднее) и второго порядка (среднеквадратичное отклонение) градиентов при оптимизации:

$$m_t = \beta_1 m_{t-1} + (1 - \beta_1) g_t$$
$$v_t = \beta_2 v_{t-1} + (1 - \beta_2) g_t^2$$

где g_t — градиент на текущей итерации, m_t — оценка первого момента, v_t — оценка второго момента, а β_1, β_2 — гиперпараметры сглаживания (обычно $\beta_1 = 0.9$, $\beta_2 = 0.999$).

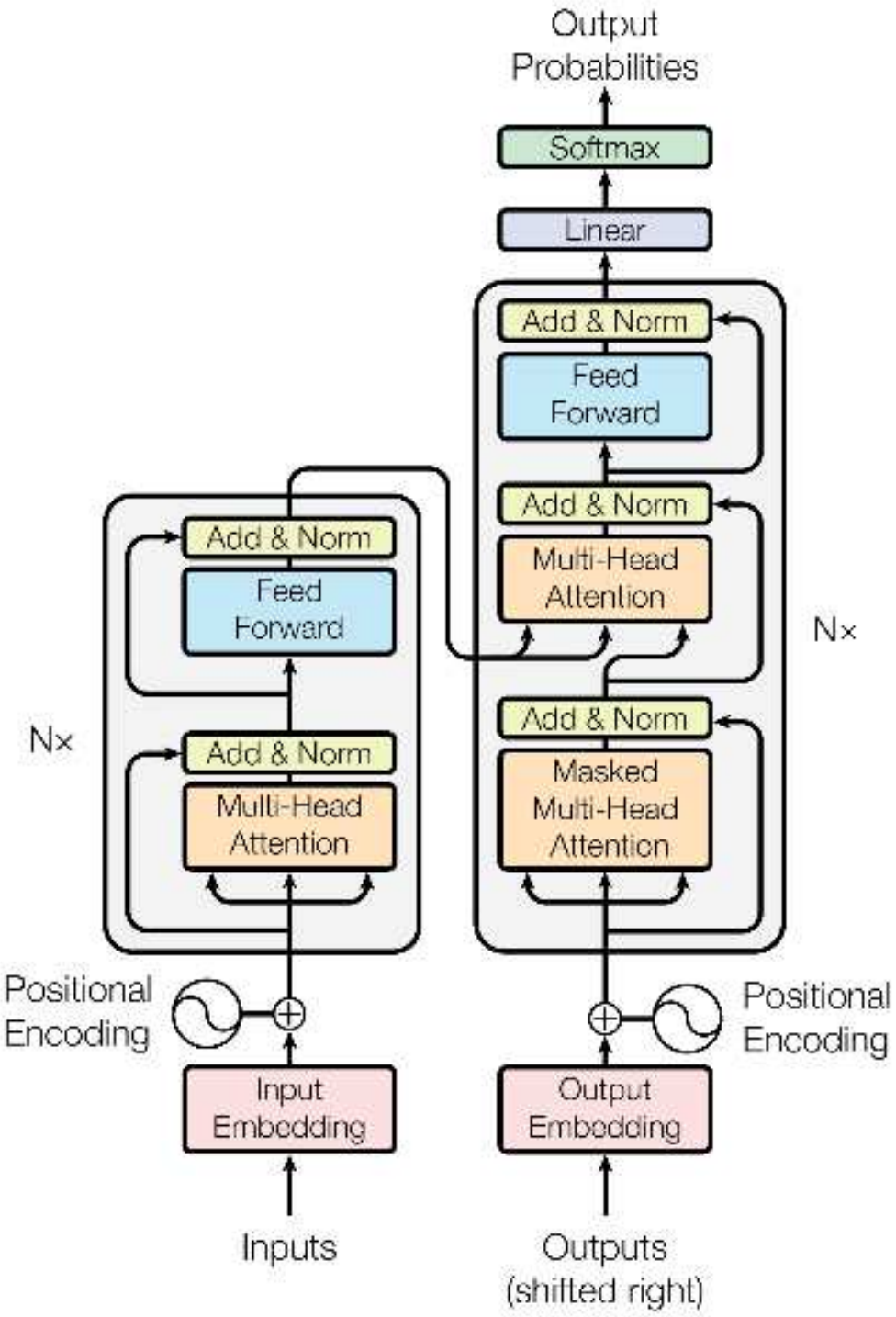


Рисунок 1.2 — Визуализация работы архитектуры Transformer.

Затем производится коррекция на смещение, возникающее в начале обучения, и обновление параметров:

$$\hat{m}_t = \frac{m_t}{1 - \beta_1^t}$$

$$\hat{v}_t = \frac{v_t}{1 - \beta_2^t}$$

$$\theta_{t+1} = \theta_t - \eta \frac{\hat{m}_t}{\sqrt{\hat{v}_t} + \varepsilon}$$

где η — базовый темп обучения, ε — малое число для избежания деления на ноль.

Adam быстро адаптируется к динамике обучающей функции, что критически важно для больших и сложных нейронных сетей, таких как Transformer. В процессе обучения весов фиксируется оптимальное значение темпа обучения и корректируются внутренние оценки дисперсии и среднего, что приводит к ускоренной и устойчивой сходимости.

Таким образом, совокупность архитектуры Transformer и алгоритма оптимизации Adam обеспечивает высокую эффективность, хорошие масштабируемость и качество генерации переводов в современных нейронных системах машинного перевода. Комбинация этих методов является де-факто стандартом индустрии и исследовательских разработок в области искусственного интеллекта для обработки естественного языка.

1.4 Оценка качества в задаче перевода

Оценка качества машинного перевода — одна из центральных и одновременно наиболее сложных задач в области автоматической обработки естественного языка. Несмотря на бурное развитие алгоритмов перевода и появление нейросетевых моделей, способных генерировать тексты, трудно отличимые от переведённых человеком, проблема объективной и надёжной оценки их качества остаётся открытой. Сложность заключается не только в семантическом богатстве естественных языков и разнообразии допустимых переводческих решений, но и в невозможности однозначно зафиксировать «идеальный» перевод для большинства предложений. Разные переводчики могут выбирать различные лексические, синтаксические или стилистические конструкции, передавая при этом одинаковый смысл, что затрудняет сравнение между реальным и автоматическим переводом посредством прямого сравнения текстов.

Из-за возможной вариативности переводов нельзя полагаться на точное совпадение с профессиональным переводом. Такими свойствами обладают метрики типа пословной точности или расстояние Левенштейна. Хорошая метрика качества должна учитывать совпадение n-грамм и несильно полагаться на порядок слов в переводе, чтобы оставить пространство для перефразирования.

В этой связи особое значение приобретают автоматические метрики, опирающиеся на различные эвристики сопоставления *n*-грамм переведённого и эталонного текстов, среди которых наибольшее распространение и признание получила метрика BLEU [18], имеющая достаточно хорошую корреляцию с оценками людей.

Метрика рассчитывается на основе пересечения *n*-грамм в автоматическом и эталонном переводах одного предложения. Для каждой *n*-граммы длины *n* рассчитывается доля P_n — среднее по всем предложениям в тестовом наборе отношения частоты *n*-граммы в кандидатах к частоте в эталонных переводах. При этом частота в кандидате ограничена значением в эталоне, чтобы отношение частот было ограничено сверху единицей:

$$P_n = \frac{\sum_{\text{n-gram} \in C} \text{Count}_{\text{clip}}(\text{n-gram})}{\sum_{\text{n-gram}' \in C'} \text{Count}(\text{n-gram}')},$$

где C - множество предложений-переводов оцениваемой модели, C' — множество эталонных переводов тестового корпуса.

Далее BLEU рассчитывается как геометрическое среднее, умноженное на константу brevity penalty (BP):

$$BP = \min(e^{1-\frac{r}{c}}, 1),$$

$$BLEU = BP \sum_{n=1}^{4} \frac{1}{n} \log P_n,$$

где r — суммарная длина эталонных переводов тестового корпуса; c — суммарная длина переводов модели. Умножение на константу BP предлагается авторами метрики для поощрения более коротких переводов системы, так как более длинные тексты в среднем содержат больше случайных пересечений по *n*-граммам с эталонными текстами.

Метрика BLEU имеет хорошую корреляцию с человеческой оценкой качества переводов как при сравнении автоматических систем с профессиональными переводчиками, так и при сравнении автоматических систем друг с другом [18]. При проведении сравнения автоматических систем перевода авторы использовали модели различной сложности, а при сравнении профессиональных переводчиков выбирались эксперты с различным уровнем владения языками.

При проведении экспериментов метрика BLEU вычисляется по тестовым корпусам, подготовленным с помощью профессиональных переводчиков. Для

экспериментов на англо-финском направлении использовались тестовые корпуса, подготовленные к конференции WMT-2017 [25]. Размер тестового корпуса для англо-финского составляет 1500 предложений, исходные предложения выбирались из новостных статей. Для оценки качества русско-английского и англо-русского переводов использовался аналогичный корпус размеров в 2000 предложений.

Глава 2. Разработка метода совместного обучение прямой и обратной моделей перевода для борьбы с нехваткой обучающих данных

В современном мире одной из ключевых проблем машинного перевода является недостаток обучающих данных. Ярким примеров этой проблемы могут служить малоресурсные языки. Для таких языков объем интернета не так велик, и автоматические корпуса, которые можно найти или построить, оказываются заметно меньшего объема [25]. При переводе на такой язык у переводной модели в обучении оказывается меньше данных для изучения лексики, грамматических конструкций и других особенностей языка. Все это негативно сказывается на итоговом качестве перевода.

При этом задача перевода на более популярный язык, например, с малоресурсного на английский, как правило, решается с лучшим качеством, чем обратная задача. Это связано с тем, что для популярных языков имеется гораздо больше разнообразных текстовых данных, например одноязычных. На них модели могут более успешно изучать особенности языка на большем объеме примеров. Таким образом, качество перевода на малоресурсные языки остаётся одной из сложных и актуальных задач машинного перевода.

В данной главе проводится обзор современных подходов к преодолению дефицита параллельных корпусов, включая методы обратного и итеративного обратного перевода, а также совместное обучение прямой и обратной моделей перевода. Описываются математические основы методов совместного обучения прямой и обратной моделей перевода, их практические и вычислительные ограничения. Приводится формализация оптимизируемых функционалов, а также даётся анализ применимости методов к современным архитектурам. Экспериментальная часть главы посвящена сравнению эффективности применяемых методов в режимах обучения с нуля и дообучения. Также приводится анализ качества моделей перевода по автоматическим метрикам и влияния синтетических данных на согласованность и точность результата.

Результаты экспериментов с использованием обратной модели для улучшения качества перевода подробно изложены в работе соискателя [55].

2.1 Обзор методов интеграции обратной модели в процесс обучения прямой модели перевода

Как уже было отмечено, при переводе с малоресурсного языка на более популярный существуют способы преодоления проблемы нехватки параллельных данных. Если для целевого языка достаточно одноязычных данных, то их можно использовать для генерации синтетических данных методом обратного перевода [6] и итеративного обратного перевода [26]. В методе обратного перевода большой объем одноязычных данных переводится с помощью обратной модели и используется для обучения прямой модели перевода. Несмотря на то, что качество таких синтетических данных может не быть идеальным, при применении данного метода модель при переводе учится воспроизводить естественные тексты на целевом языке, разнообразие которых может быть больше чем внутри параллельных данных [6]. За счет этого при наличии большого количества одноязычных данных можно добиться улучшения качества перевода при недостатке параллельных данных для обучения. Стоит отметить, что в случае перевода на малоресурсный язык данный метод может оказаться менее эффективным, так как для такого языка количество одноязычных данных также невелико.

Метод итеративного обратного перевода отличается от метода обратного тем, что процедура генерации синтетических данных повторяется заново после каждой итерации обновления модели. За счет этого качество синтетических данных с каждой итерацией становится лучше, что приводит к улучшению качества модели перевода после переобучения.

Как было показано выше, традиционные методы обучения моделей перевода могут оказаться недостаточно эффективными в условиях недостатка обучающих данных. В связи с этим возникает необходимость в разработке и применении новых подходов, которые позволят улучшить качество перевода даже при ограниченном объёме данных. Одним из перспективных направлений является совместное обучение прямой и обратной моделей перевода. Совместное обучение позволяет использовать информацию, содержащуюся в обратной модели, для улучшения качества прямой модели и наоборот. Это особенно актуально для языковых пар с дефицитом параллельных данных, поскольку такой подход способствует более эффективному использованию имеющихся ресурсов и повышению согласованности переводов.

В работе [7] авторами был предложен метод совместного обучения прямой и обратной моделей перевода. Для этого предлагается учить модель генерировать такие переводы, чтобы перевод обратной модели больше совпадал с исходным текстом. При этом авторы обучают такую модель с помощью алгоритма REINFORCE с эмпирически выбранной функцией награды. В результате применения алгоритма градиент для прямой модели можно записать следующим образом:

$$
\begin{aligned}
&y \sim P_\Theta(y|x) \\
&\nabla_\Theta L(x) = \big(\alpha P_{LM}(y) + (1-\alpha)\log P'(x|y)\big)\nabla_\Theta \log P_\Theta(y|x),
\end{aligned}
\tag{2.1}
$$

где $P_\Theta(y|x)$ — параметризованная прямая модель перевода, $P'(x|y)$ — обратная модель, $P_{LM}(y)$ — дополнительная фиксированная языковая модель, а α — гиперпараметр, подбираемый экспирически.

Оказалось, что такой метод увеличивает согласованность и качество перевода в целом за счет предоставления прямой модели во время обучения информации, которую несет в себе обратная модель.

Однако применение описанного метода сопряжено с некоторыми трудностями. Одна из них заключается в необходимости одновременно хранить градиенты для прямой и обратной моделей на каждой итерации обучения. В предложенном авторами методе предполагается использовать рекуррентные модели перевода [1], но для современных архитектур такие требования делают метод труднопримениым на практике, так как гиперпараметры прямой модели для наиболее эффективного обучения выбираются так, чтобы использовать всю возможную память вычислительных устройств [27]. Другая проблема описанного подхода заключается в том, что на первых итерациях обучения градиент для прямой модели обладает высокой дисперсией. По этой причине в [7] для стабилизации обучения оценивается значимость обратного перевода с помощью дополнительных языковых моделей, что требует еще больше вычислительных ресурсов и времени. Из-за всего этого данный метод труднопримениим для обучения больших моделей перевода в современных реалиях. Более того, эмпирический выбор функции потерь для обучения оставляет простор для интерпретации и улучшения предложенного авторами метода обучения.

В данной работе соискателя представлен новый метод обучения прямой модели перевода с интеграцией обратной модели в процесс обучения. Алгоритм

обучения предложенного соискателем метода выведен из задачи максимизация правдоподобия циклических переводов и является теоретически и практически обоснованным. Полученная формула функции потерь отличается от формулы в [7]. В ней удалось избежать эвристики с использованием дополнительных языковых моделей для стабилизации обучения. Проблема высокой дисперсии градиентов, даваемых полученной в исследовании соискателя функцией потерь, решается с помощью дообучения уже предобученной модели перевода. Данное решение экономит вычислительные ресурсы, так как нет необходимости хранить в памяти дополнительные параметры языковых моделей и статистики оптимизатора для обратной модели, а сам процесс дообучения сходится существенно быстрее, чем обучение с нуля. Все это позволяет обучать прямую модель совместно с обратной, используя современные архитектуры моделей перевода, не уменьшая их в размере.

Результаты исследования могут найти практическое применение в области машинного перевода для малоресурсных языков. Предложенные соискателем методы способствуют развитию методов обучения моделей в условиях ограниченного объёма данных.

2.2 Метод максимального правдоподобия для циклического перевода

В данной работе ростом согласованности моделей будем считать увеличение числа совпадений циклического перевода через целевой язык с оригинальным текстом. Для этого рассмотрим конструкцию порождения циклического перевода ЯВ$\rightarrow$ ЦЯ$\rightarrow$ ЯВ. В нем текст на языке входа прямой моделью переводится на целевой язык, а потом переводится обратной моделью снова на язык входа. Для построения вероятностной модели необходимо задать требование совпадения полученного циклического перевода с исходным текстом.

Для этого определим понятие вероятности циклического перевода следующим образом:

$$P_{\text{cycle}}(x'|x) = \mathbb{E}_{y\sim P_\Theta(y|x)} P'(x'|y), \tag{2.2}$$

где $P_\Theta(y|x)$ — параметризованная прямая модель перевода, а $P'(x|y)$ — обратная модель. Такая вероятность показывает то, какова в среднем по всем прямым переводам модели вероятность получения циклического перевода x' из текста x.

Теперь, когда получена вероятность циклического перевода $P_{\text{cycle}}(x'|x)$, запишем оптимизируемый функционал который отражает необходимость совпадения циклического перевода с исходным текстом:

$$\log P_{\text{cycle}}(x|x) = \log \mathbb{E}_{y \sim P_\Theta(y|x)} P'(x|y) \longrightarrow \max_\Theta . \tag{2.3}$$

Данный функционал можно рассматривать как метод максимизации правдоподобия для циклического перевода 2.2.

Рассмотренный подход построения и оптимизации вероятности совпадения циклического перевода концептуально схож с методологией автокодировщика, распространённой в задачах обучения представлений. В классическом автокодировщике входной объект последовательно преобразуется кодировщиком в скрытое представление, а затем декодировщиком реконструируется обратно в исходное пространство. Качество автокодировщика измеряется степенью близости восстановленного объекта к исходному входу, что стимулирует модель учиться компактно и содержательно кодировать ключевую информацию о входных данных.

Аналогично, в задаче машинного перевода с использованием циклической функции — сначала исходный текст на языке входа переводится прямой моделью на целевой язык, а затем полученный перевод обратной моделью снова переводится обратно на язык входа. Совпадение «циклического» и исходного текста становится метрикой согласованности двух моделей, что побуждает их не только учиться точному переводу в одну сторону, но и поддерживать обратимость переводческого процесса. Такой подход обладает важными преимуществами: он способствует обучению моделей переводить не только отдельные устойчивые выражения, но и улавливать глобальные смысловые и структурные зависимости между языками, а также использовать знания обеих языковых систем для повышения надежности перевода. Кроме того, как и в автокодировщике, подобная структура помогает выявлять и минимизировать скрытые ошибки, такие как искажения, пропуски и избыточности, что в конечном счёте повышает как качество итогового перевода, так и степень взаимной согласованности прямой и обратной моделей.

Несмотря на полезность введенного понятия 2.2 для построения вероятностной модели, вычислить его невозможно, так как подсчет среднего по всем возможным переводам приводит к суммированию по бесконечному множеству текстов на целевом языке. Для получения несмещенной оценки этого выражения, которое понадобится для оценки функции потерь на тестовой выборке, необходимо пользоваться процедурой выбора по значимости.

Для обучения нейросетевых моделей методом градиентного спуска получим значения градиента данной функции потерь. После произведенного логарифмирования выражение более невозможно оценить несмещенно с помощью процедуры выбора по значимости, поэтому перейдем к максимизации нижней оценки функции потерь, воспользовавшись при этом неравенством Йенсена для вогнутых функций. Функция логарифма является вогнутой, а вероятностное распределение образует выпуклую комбинацию, следовательно неравенство Йенсена применимо:

$$\log \mathbb{E}_{y\sim P_\Theta(y|x)} P'(x|y) \geqslant \mathbb{E}_{y\sim P_\Theta(y|x)} \log P'(x|y) =: L(x).$$

После получения нижней оценки запишем для нее значение градиента. Представим математическое ожидание в виде интеграла и внесем градиент, являющийся линейным оператором, внутрь интеграла:

$$\begin{aligned}\nabla_\Theta L(x) = \nabla_\Theta \int P_\Theta(y|x) \log P'(x|y) dy = \\ = \int \log P'(x|y) \nabla_\Theta P_\Theta(y|x) dy.\end{aligned}$$

Чтобы избежать суммирования по бесконечному множеству, при подсчете градиента также воспользуемся процедурой выбора по значимости. Для этого данное представление градиента нижней оценки функции потерь представим в виде математического ожидания по некоторому распределению, воспользовавшись следующим преобразованием:

$$\frac{\partial \log f(x)}{\partial x} = \frac{1}{f(x)} \frac{\partial f(x)}{\partial x}$$

Применив данное преобразование, получим следующий вид градиента нижней оценки функции потерь:

$$\begin{aligned}\nabla_\Theta L(x) = \int \log P'(x|y) P_\Theta(y|x) \nabla_\Theta \log P_\Theta(y|x) dy = \\ = \mathbb{E}_{y\sim P_\Theta(y|x)} \log P'(x|y) \nabla_\Theta \log P_\Theta(y|x)\end{aligned}$$

Теперь есть возможность оценить данное выражение значением на одном примере с помощью процедуры выбора по значимости:

$$\nabla_\Theta L(x) \approx \log P'(x|y) \nabla_\Theta \log P_\Theta(y|x), \quad y \sim P_\Theta(y|x). \tag{2.4}$$

Если рассмотреть $\mathcal{L}(x, y)$ — функцию потерь прямой модели при обучении без циклических переводов 1.1, то полученная оценка градиента 2.4 может быть представлена как градиент при обучении без циклических переводов, умноженный на некоторый коэффициент:

$$\nabla_\Theta \mathcal{L}(x, y) = \nabla_\Theta \log P_\Theta(y|x)$$
$$\nabla_\Theta L(x) \approx w(x, y) \nabla_\Theta \log P_\Theta(y|x), \quad y \sim P_\Theta(y|x),\ w = \log P'(x|y).$$

Представление градиента функции потерь в виде градиента прямой модели, умноженной на коэффициент, позволяет дать интерпретацию полученным формулам: чем больше вероятность получения исходного текста при обратном переводе из y, тем более прямая модель будет стремиться переводить x как y в сравнении с обучением без циклической функции потерь 1.1. Данная интерпретация подтверждает выдвинутое выше предположение, что оптимизация данного функционала аналогично автокодировщику толкает модель при переводе сохранять в переводе весь заложенный в исходном тексте смысл так, чтобы по нему можно было восстановить исходный текст.

При полученном визуальном сходстве градиентов функций потерь прямой модели перевода 1.1 и модели с циклическими переводами 2.3 в последней вместо реальных текстов используются сгенерированные с помощью прямой модели переводы.

Можно заметить, что представленное выражение градиента 2.4 имеет сходство с выражениями градиентов из работы [7], представленными в формуле 2.1, при том, что ее авторы получили градиенты для обучения, используя эвристики. В настоящем исследовании соискателем представлена исходная вероятностная модель, из которой получены градиенты для оптимизации. Данная вероятностная модель обобщает использование обратной модели перевода и дает возможность для развития данного метода в будущем.

2.3 Оценка качества циклического перевода

В данной работе в рамках улучшения качества перевода предлагается улучшение согласованности переводов прямой модели и обратной ей. Для оценки улучшения согласованности требуется оценить, насколько сильно циклический перевод исходного текста искажает этот текст. С этим может помочь уже описанная метрика BLEU [18], позволяющая вычислять сходство между двумя текстами на одном языке. В данной работе метрику BLEU, рассчитанную для исходных текстов и их циклических переводов, будем называть CycleBLEU, и с ее помощью будем оценивать рост согласованности переводов прямой и обратной ей моделей.

В экспериментах ожидается, что после обучения с циклической функцией потерь 2.3 будет наблюдаться рост согласованности переводов, который можно будет увидеть с помощью метрики CycleBLEU, даже если прироста BLEU наблюдаться не будет. Это мотивируется тем, что рост согласованности прямой и обратной ей моделей сам по себе является достижением, даже при отсутствии заметного улучшения качества перевода на тестовых наборах.

2.4 Эксперименты с совместным обучением прямой и обратной моделей

Далее подробно описываются условия проведения экспериментов, выбор архитектурных решений и полученные результаты для моделей перевода, обученных с использованием циклической функции потерь 2.3. Все нейросетевые переводчики, рассматриваемые в экспериментах, были построены на базе архитектуры Transformer, что обусловлено её доказанной эффективностью в задаче машинного перевода и возможностью масштабирования на большие объемы данных.

2.4.1 Детали выбранных архитектур

Экспериментальная часть была условно разделена на два направления: первое — исследование обучения моделей с нуля с циклической функцией потерь 2.3, второе — изучение дообучения предобученных моделей с той же функцией потерь. Для первой серии экспериментов использовалась компактная конфигурация Transformer-tiny, в которой размер промежуточных векторных представлений составлял 256, а внутренние размерности Feed-Forward Network (FFN) блоков — 1024. Выбор небольшой модели продиктован задачей анализа особенностей сходимости и устойчивости обучения в условиях высокой дисперсии градиентов; в этих экспериментах целью было не достижение финального высокого качества, а исследование поведения модели и закономерностей оптимизационного процесса при разных настройках.

Для серии экспериментов с дообучением на циклическую функцию потерь 2.3 в качестве исходных использовались модели перевода, предварительно обученные на классической функции без циклических переводов 1.1. В качестве архитектуры была выбрана конфигурация Transformer-base. Здесь размерность промежуточных представлений составляла уже 512, а внутри FFN-блоков она увеличивалась до 2048, что соответствует промышленным стандартам современных моделей нейронного машинного перевода и обеспечивает значительный рост потенциального качества итоговой модели.

Во всех проведённых экспериментах для повышения качества и устранения эффекта переобучения под конкретные позиции внутри предложений применялся механизм относительного кодирования позиций [28], интегрированный в оригинальную архитектуру Transformer. Такой подход доказал свою полезность при обучении на длинных и разнообразных по структуре последовательностях, делая модель более гибкой и универсальной.

Процесс оптимизации параметров всех сетей происходил с использованием стохастического оптимизатора Adam, что является индустриальным стандартом для задач глубокого обучения. Для дообучения моделей использовался специальный режим прогрева (warm-up): в течение первых 2000 итераций шаг градиента постепенно увеличивался для более надёжного накопления статистик первых и вторых моментов, необходимых этому алгоритму для устойчивой и быстрой сходимости. После этапа нагрева базовый шаг обучения фиксировался на уровне 10^{-5}

с последующим уменьшением по мере приближения к минимуму функции потерь.

Все вычислительные эксперименты осуществлялись на вычислительном кластере, оснащённом восемью графическими ускорителями Tesla M40. Для эффективного использования вычислительных ресурсов обучающий батч данных на каждой итерации равномерно распределялся между всеми GPU. Параллельные вычисления градиентов позволяли существенно ускорить процесс обучения.

2.4.2 Данные для экспериментов

В рамках данного исследования эксперименты проводились на двух относительно малоресурсных языковых направлениях: англо-финском и русско-казахском. Для обучения моделей на обоих направлениях использовались корпуса, полученные путём автоматизированного сбора данных из интернета. Подготовка параллельных данных по мультиязычным данным из интернета — достаточно трудоемкий процесс, состоящий из нескольких этапов: парсинг HTML страниц [29], построение кандидатов на параллельные документы, фильтрация пар документов и выравнивание переведенных предложений [30], [31], [32]. В данном исследовании в качестве основы для параллельных данных выступали интернет-страницы, имеющие переведённые версии. В итоге для каждого языкового направления было собрано порядка 50 миллионов пар параллельных предложений.

Так как в основе собранных корпусов для обоих направлений лежат интернет-страницы, имеющие версии на разных языках, в корпусе присутствует тематическое смещения. Например, тематика энциклопедических статей и новостей представлена значительно выше, чем в случайной одноязычной выборке документов на каждом из целевых языков, а тематика научных текстов — ниже. Средняя длина одного предложения на финском в англо-финском параллельном корпусе составила 10.3 слова. Средняя длина одного предложения на казахском в русско-казахском корпусе составила 13.4 слова.

Для предварительной обработки текстов всех экспериментов применялось разбиение слов на подслова [2]. BPE уменьшает общее количество уникальных токенов, присутствующих в обучающей выборке, разбивая редкие и составные

слова на более частотные элементы. Это позволяет существенно снизить размер итогового словаря (до 32 000 токенов для каждой модели), повысить устойчивость модели к неизвестным словам и ускоряет процесс обучения за счёт сокращения числа параметров.

Помимо параллельных пар предложений в обучении широко использовались синтетические пары. Они были получены путём обратного перевода: модели, обученные переводить в противоположном направлении (например, с финского на английский), использовались для перевода больших массивов одноязычных текстов целевого языка на язык источника. Для получения синтетических переводов использовались обратные модели аналогичного размера, обученные на тех же параллельных данных. В обучении обратных моделей также присутствовали синтетические обратные переводы, которые были получены с помощью прямой модели, но обученной уже без синтетических данных.

Такие синтетические примеры затем добавлялись в обучающую выборку для основной модели. Следует отметить, что значительное присутствие синтетических данных в обучении может уменьшить дополнительный эффект от применения циклической функции потерь 2.3, поскольку информация от обратной модели уже частично интегрируется в обучение через сами синтетические переводы. Этот возможный эффект дополнительно исследуется в экспериментальной части.

Выбор именно этих языковых пар обусловлен тем, что на данных направлениях объём качественных текстов на целевом языке в открытом доступе относительно невелик, что типично для малоресурсных языков. К примеру, в случае русско-казахского синтетические пары были составлены на основе 5 миллионов предложений из News Crawl 2018, что в десять раз меньше размера параллельного корпуса. Для англо-финского направление качество и масштаб синтетических данных сопоставимы с параллельным корпусом: для этого использовались 50 миллионов предложений из News Crawl 2014. Следует подчеркнуть, что при составлении синтетических корпусов на целевом языке особое внимание уделялось качеству текстов, так как наличие ошибок или сбоев в таких данных может отрицательно сказаться на итоговой производительности переводных моделей. В результате построения и отбора данных удалось обеспечить надёжную экспериментальную основу для анализа эффективности рассматриваемых алгоритмов и стратегий обучения.

2.4.3 Эксперименты с обучением с нуля

При обучении с циклической функцией потерь 2.3 удалось получить формулы для вычисления градиентов, которые имеют общее с формулами градиентов из оригинальной статьи [7]. Однако важным отличием является то, что в предложенной функции потерь 2.3 не используются вспомогательные языковые модели. Хранение языковых моделей в памяти вычислительных устройств существенно ограничит количество свободных ресурсов, что приведет к уменьшению реального количества данных, обрабатываемых на одной итерации обучения. Уменьшение количества обрабатываемых данных на устройстве неизбежно приведет к заметному падению качества.

Кроме того, чтобы уменьшить расход памяти вычислительного устройства, будем оптимизировать только параметры прямой модели с помощью градиентов от циклических переводов 2.3, а в качестве обратной модели будем использовать уже предобученную. В описанных условиях обучение модели архитектуры Transformer-base с нуля составляет чуть менее двух недель.

Описанные ограничения вычислительных ресурсов делают применение оригинального метода [7] неэффективным для современных нейросетевых архитектур. Для хранения в памяти видеокарты состояния оптимизатора обратной модели, весов вспомогательных языковых моделей потребовало бы уменьшения размера прямой модели перевода в несколько раз. Всё это делает метод из статьи [7] применимым только к устаревшим архитектурам моделей на практике. Поэтому предложенный соискателем метод исследуется отдельно.

Важнейшей задачей является анализ сходимости процесса обучения модели перевода с применением циклической функции потерь 2.3 в актуальной архитектуре Transformer. Для получения наиболее наглядных и интерпретируемых результатов на первом этапе исследования выбиралась компактная конфигурация модели — Transformer-tiny, содержащая существенно меньшее количество параметров по сравнению с Transformer-base.

Обучение модели осуществлялось на направлении с английского на финский язык, используя предварительно подготовленный корпус параллельных данных в сочетании с добавлением синтетических примеров. В общей сложности обучение велось на протяжении 30 000 итераций.

Результаты сравнения качества достигнутых моделей, представленные в таблице 1, показывают, что использование циклической функции потерь 2.3 на практике заметно уступает по итоговой метрике BLEU обучению модели с функцией потерь 1.1. По полученным результатам видно, что обучение с использованием обратной модели перевода уступает обучению только прямой модели по метрике BLEU на 2 пункта и по метрике CycleBLEU на 8 пунктов. Результаты данной серии экспериментов приведены только после настройки оптимальных гиперпараметров обучения, то есть представляют наиболее благоприятные условия для каждой из схем. Отдельно следует отметить, что значение метрики CycleBLEU при обучении с циклической функцией потерь оказалось достаточно низким, несмотря на то, что теоретически предложенный соискателем подход напрямую должен стимулировать ее рост.

Таблица 1 — Результаты при обучении с нуля на англо-финском направлении с использованием обратной модели перевода и без.

Модель	BLEU	CycleBLEU
Прямая	12.0	50.0
Прямая+обратная	10.0	42.0

Данные результаты объясняются тем, что на первых этапах обучения градиенты, даваемые циклической функцией потерь 2.4, обладают крайне высокой дисперсией из-за того, что переводы, генерируемые прямой моделью, обладают низким качеством. Для таких переводов оценка обратной моделью является крайне шумной и не несет полезной информации для обучения. Решить данную проблему можно с использованием предобученной модели перевода для инициализации прямой модели в процедуре обучения с циклической функцией потерь 2.3. В таком случае переводы, генерируемые прямой моделью, с самого начала обучения будут качественными.

2.4.4 Эксперименты с дообучением моделей

Для инициализации прямой модели в экспериментах по дообучению использовалась модель архитектуры Transformer-base, предварительно обученная

традиционным образом до сходимости по функции потерь без циклических переводов 1.1. Это позволило обеспечить высокое базовое качество начального состояния модели и, таким образом, сфокусироваться на влиянии самого процесса дообучения с использованием циклической функции потерь.

Данные, применяемые на этапе дообучения, полностью совпадали с теми, что использовались при исходном обучении: они включали как параллельные пары предложений, так и синтетически сгенерированные примеры, полученные методом обратного перевода. Благодаря тому, что выбранный шаг градиентного обновления весов в процессе дообучения был существенно снижен по сравнению с этапом предобучения, наблюдалась тенденция к продолжению роста качества модели даже за счет простого продолжения обучения без изменения функции потерь. В связи с этим для объективной оценки эффективности метода дообучения с циклической функцией потерь 2.3 полученные результаты сопоставлялись как с исходной (предобученной) моделью, так и с моделью, дополнительно дообучавшейся по классической стратегии без циклических переводов 1.1.

Анализ результатов, приведённых в таблице 2 для направления перевода с английского на финский, показывает, что дообучение с введением циклической функции потерь обеспечивает прирост по метрике качества перевода BLEU на 0.5-2 пункта по сравнению с обоими контрольными случаями. Прирост в 0.5 пункта значим, хоть и в соответствии с некоторыми исследованиями метрики BLEU, может не приводить к однозначному улучшению с точки зрения человеческой разметки [33], [34]. Это позволяет сделать вывод, что интеграция циклического компонента в процесс дообучения способствует повышению качества модели перевода с точки зрения метрики BLEU, даже при использовании обратных переводов в обучении модели перевода.

Таблица 2 — Результаты дообучения с циклической функцией потерь для англо-финского направления

Модель	BLEU	CycleBLEU
Без дообучения	23.4	47.0
Прямая	25.0	55.0
Прямая+обратная	25.5	54.5
Google Translate (2024)	25.9	-

Полученные результаты дообучения с использованием циклической функции потерь 2.3 оказываются в определённой степени парадоксальными, если

их анализировать с точки зрения метрики CycleBLEU. Несмотря на то что предложенный соискателем метод специально ориентирован на максимизацию согласованности моделей в циклической задаче, итоговое значение CycleBLEU оказалось ниже по сравнению с обычным дообучением 1.1 на 0.5 пункта. На первый взгляд, это противоречит интуитивным ожиданиям, поскольку циклическая функция потерь должна была бы непосредственно способствовать улучшению совпадения между исходными и циклически восстановленными текстами.

Тем не менее, данная особенность объясняется значительной долей синтетических примеров в обучающей выборке, полученных посредством обратного перевода. Поскольку при формировании этих данных обратная модель уже оказывает прямое влияние на обучение прямой модели, информация о структуре и свойствах целевого языка, а также о возможных ошибках, недостаточностях и вариантах переводов, фактически встраивается в процесс обучения еще до введения циклической функции потерь. Таким образом, для рассматриваемой экспериментальной конфигурации оказывается, что использование синтетических данных само по себе может быть достаточным для достижения высокого уровня согласованности циклических переводов, и дополнительный вклад от оптимизации по циклической функции потерь становится менее заметным. Тем не менее, само качество перевода, измеряемое метрикой BLEU, показывает рост качества перевода, что также может объясняться тем, что метрика CycleBLEU не является оптимизируемым функционалом и может недостаточно оценивать все аспекты сохранения смысла в задаче перевода.

Рассмотрим теперь эксперименты с дообучением для направления с русского на казахский. Результаты эксперимента приведены в таблице 3. В данном эксперименте также наблюдается рост метрики BLEU относительно модели без дообучения, однако такой прирост достаточно мал и не является значимым [33], [34]. Однако прирост метрики CycleBLEU на 2.4 пункта значим и говорит о росте согласованности переводов прямой и обратной моделей друг с другом.

Сравнение с качеством перевода Google Translate для обоих направлений перевода приведено для ориентира. Выводы на основе этого сравнения сделать затруднительно, так как неизвестны характеристики данной системы, а качество системы меняется со временем.

Анализируя полученные результаты на русско-казахском направлении, можно отметить, что именно здесь наблюдается значительный прирост значения метрики CycleBLEU при обучении с циклической функцией потерь 2.3.

Таблица 3 — Результаты дообучения с циклической функцией потерь для русско-казахского направления

Модель	BLEU	CycleBLEU
Без дообучения	19.5	41.5
Прямая	19.80	42.5
Прямая+обратная	20.05	44.9
Google Translate (2025)	17.9	-

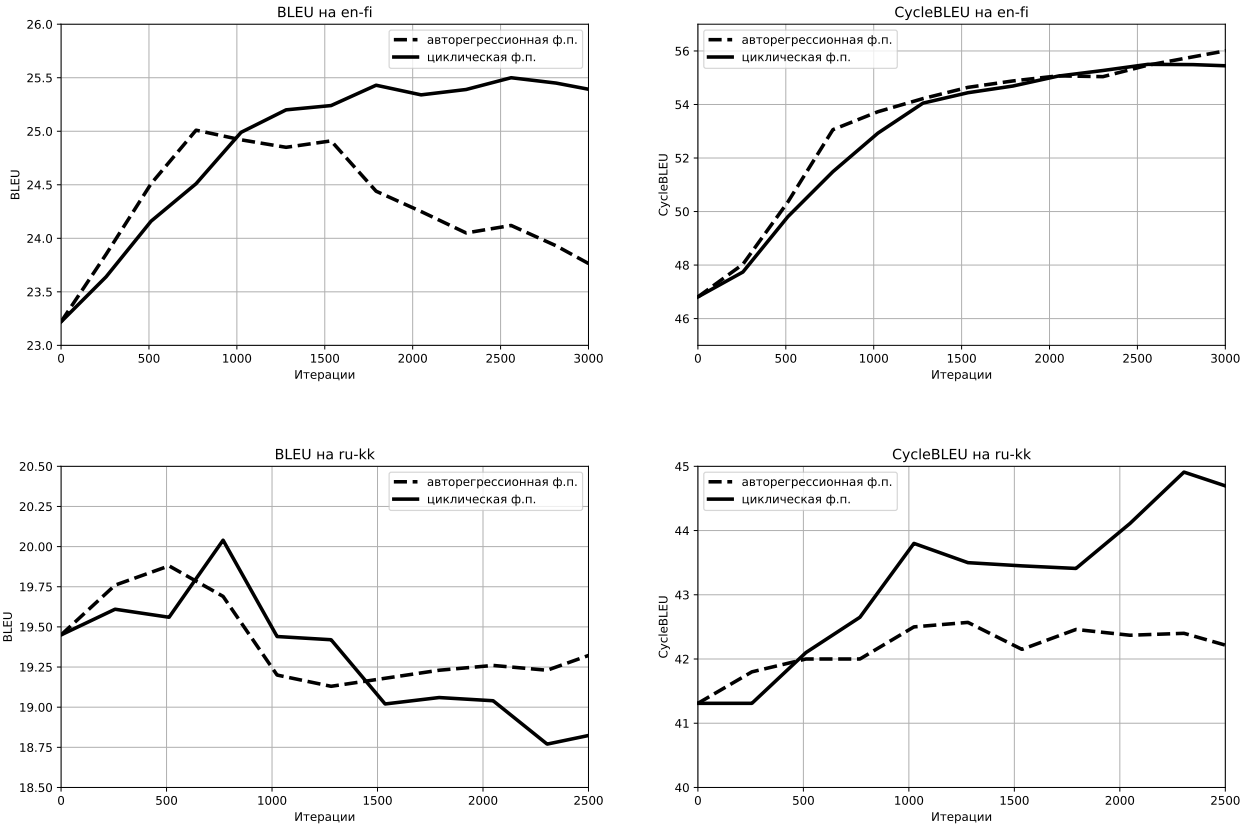
Это контрастирует с более скромными изменениями CycleBLEU в случае англо-финского направления и указывает на особую роль соотношения синтетических и параллельных данных в обучающей выборке. На русско-казахском корпусе доля синтетических примеров заметно ниже, чем на англо-финском, вследствие чего эффект от добавления специальной циклической функции потерь становится более выраженным.

Данная картина позволяет сделать вывод, что использование либо синтетических данных, полученных обратным машинным переводом, либо оптимизация по циклической функции потерь приводят к схожим результатам с точки зрения метрики совпадения исходных и циклических переводов (CycleBLEU). В ситуации, когда обучение модели уже включает значительное число синтетических примеров, дополнительное дообучение на циклической функции потерь практически не повышает значения CycleBLEU. Напротив, если синтетических данных мало, именно циклическая оптимизация позволяет существенно повысить согласованность между прямой и обратной моделями.

Динамика обеих метрик в ходе обучения детально представлена на графике 2.1: для англо-финского направления особенно чётко проявляется рост BLEU при низком уровне колебаний метрик, однако CycleBLEU практически не улучшается. В то же время на русско-казахском направлении кривая CycleBLEU демонстрирует явное увеличение, сопровождающееся относительно большим шумом значений, что объясняется небольшим объёмом тестового корпуса и повышенной долей случайных вариаций при его оценке. Это дополнительно подчеркивает важность как структуры обучающих данных, так и характеристик используемых метрик при интерпретации итоговой производительности моделей перевода в режиме совместного обучения.

В итоге общее число итераций, во время которых качество моделей улучшается при дообучении, не превосходит 3000, что в 10 раз меньше количества

Рисунок 2.1 — Графики дообучения моделей с использованием циклической функции потерь и без. Верхняя строка соответствует направлению en-fi, нижняя — направление ru-kk. Первая колонка соответствует графикам BLEU, вторая — графикам CycleBLEU.



итераций, необходимых для обучения модели перевода с нуля. Таким образом, улучшения качества при использовании функции потерь с циклическими переводами 2.3 удается добиться при более быстром обучении, чем при обучении с нуля, как в оригинальной работе [7].

2.5 Выводы

В целом, использование дообучения с циклическими переводами 2.3 позволяет улучшить качество переводов, хотя прирост оказался меньше, чем заявлялся в оригинальной работе [7]. Это может объясняться более высоким качеством, даваемым современными архитектурами в задаче машинного перевода. При этом отдельно стоит отметить вычислительную эффективность предложенной

соискателем процедуры дообучения, которая позволяет применить совместное обучение с обратной моделью в моделях перевода с современными архитектурами Transformer.

Получен теоретический вывод метода совместного обучения из задачи максимизации циклического правдоподобия, что отличает подход от статьи [7], где формулы получены эвристически. Также предложенный подход позволяет не использовать вспомогательных языковых моделей, что облегчает его применение на практике.

Дообучение с введением циклической функции потерь обеспечивает прирост по метрике качества перевода BLEU на 0.5 пункта по сравнению с дообучением без циклической функции потерь на англо-финском направлении перевода и 0.25 — на русско-казахском направлении. На русско-казахском направлении также наблюдается значимый рост метрики CycleBLEU, что говорит о росте согласованности переводов прямой и обратной модели.

Таким образом, метод совместного обучения с использованием обратной модели можно считать эффективным с точки зрения метрики качества BLEU для использования в обучении малоресурсных направлений, а также его применение может повышать согласованность переводов прямой и обратной модели. Однако размер прироста зависит от направления перевода, количества обучающих данных и доли синтетических обратных переводов в обучении.

Глава 3. Разработка метода переупорядочивания гипотез с помощью разметки, выполненной человеком, для борьбы с систематическими ошибками перевода

Одна из ключевых проблем современных систем машинного перевода — наличие систематических ошибок, связанных с ограниченностью и неидеальным качеством параллельных данных. Для обучения нейросетевых алгоритмов, лежащих в основе таких систем, требуется большое количество выравненных пар текстов на разных языках. Однако в силу высокой стоимости профессионального перевода и сложности автоматического сбора таких корпусов, большая часть используемых данных собирается автоматически с помощью эвристических методов [32]. Это зачастую приводит к недостаточной выравненности и качеству обучающих данных, что влияет на результаты перевода. В частности, одной из типовых проблем становятся недостаточные, или неполные, переводы, а также искажения и упущения смысловых элементов [8].

Для борьбы с систематическими ошибками существует множество подходов. Так, в работе [35] авторы обратили внимание на гендерные сдвиги в перевододных моделях и предложили метод добавления гендерного сигнала в системы машинного перевода. В работе [36] авторы отмечают проблему дословного перевода (translationese), свойственную системам перевода, и предлагают метод улучшения качества на основе добавления валидированных качественных переводов в обучение и применения специальных методов фильтрации. Предложенные подходы помогают уменьшить количество конкретных типов ошибок при переводе, но требуют усложнения процесса обучения и не переносятся на другие типы систематических ошибок.

Также для борьбы с недостаточными и избыточными переводами авторами [8] был предложен контрастный подход, который заключается в интеграции негативных примеров в процесс обучения модели. Его суть состоит в том, чтобы обучать модель различать качественные и ошибочные переводы. На практике негативные примеры обычно создаются искусственно — например, путём удаления случайных слов из правильных переводов. Это позволяет повысить внимательность модели к деталям оригинального текста и снижает количество ошибок, связанных с пропусками или искажением смысла. Тем не менее, такой подход имеет ограничения: негативные примеры, сгенерированные по фиксированным шаблонам, охватывают лишь ограниченный набор ошибок и не всегда

представляют сложные ситуации, с которыми модель сталкивается на практике. Кроме того, для каждого типа ошибок требуется отдельная логика генерации негативных примеров, что снижает универсальность и масштабируемость метода.

В данной работе соискателем предлагается подход, позволяющий перейти от шаблонной генерации негативных примеров к более естественной их формулировке на основе самой работы модели. Для каждого входного текста с помощью разнообразного поиска или сэмплирования с температурой строится набор возможных переводов, среди которых с помощью оценок, выполненных людьми, определяется лучший (позитивный пример), а остальные рассматриваются как негативные относительно него. Такой подход позволяет формировать действительно сложные негативные примеры, поскольку они являются результатом реальных ошибочных предсказаний модели. Далее проводится обучение на упорядочивание этих гипотез в соответствии с их оценкой качества, проведенной людьми. Достоинством метода является отсутствие необходимости в эталонных переводах или параллельных данных, что делает его особенно полезным для специфических доменов, для которых недостаточно разметки для классической доменной адаптации. В работе рассматривается применение переупорядочивания гипотез с помощью разметки к задаче перевода товарных заголовков в сфере электронной коммерции, где тексты обладают специализированной лексикой и структурой и, как следствие, представляют особую сложность для автоматического перевода.

Экспериментальные результаты показывают, что обучение с упорядочиванием гипотез по оценкам, выполненным людьми, позволяет не только повысить общее качество перевода, но и существенно снизить распространённость наиболее типовых ошибок, делая применение нейросетевого машинного перевода более надёжным для практических задач в узких доменах.

Основные новшества предложенного соискателем подхода перечислены ниже.

1. **Предложенный метод позволяет автоматизировать получение негативных примеров из гипотез самой модели.** Для каждого входного текста с помощью процедур разнообразного поиска (например, diverse beam search) или сэмплирования с температурой формируется набор возможных переводов. Это позволяет получать негативные примеры, естественным образом отражающие ошибки, которые модель реально совершает при генерации переводов.

2. **Использование человеческих оценок для ранжирования переводов** Из набора переводов выбирается наилучший вариант на основании оценки аннотаторов, который рассматривается как позитивный пример, тогда как остальные, менее удачные варианты — как негативные относительно него. Такой подход позволяет делать критику модели более точечной и учитывать именно те ошибки, которые воспринимаются как значимые профессиональными перефодчиками.
3. **Независимость от наличия эталонного перевода**. Метод не требует параллельных данных, что позволяет получать улучшение в качестве при отсутствии подготовленных корпусов переводов. Это открывает возможность улучшения качества перевода в сложных доменах, в которых данные отсутствуют или их мало, например в электронной коммерции.

Подробные результаты исследования метода переупорядочивания гипотез изложены в работе диссертанта [54]. Применение данного подхода к задаче дообучения языковой модели под задачу перевода изложены в [59], в которой диссертант отвечал за эксперименты по интеграции размеченных людьми предложений в адаптированную к задаче перевода языковую модель.

3.1 Контрастное обучение с негативными примерами

Контрастное обучение с негативными примерами позволяет целенаправленно повысить чувствительность модели к различным видам ошибок, которые часто возникают из-за недостаточно точного соответствия между исходным и переводным текстом в параллельных данных. Суть этого подхода заключается в том, чтобы обучить модель не только максимально увеличивать вероятность правильного перевода, но и одновременно минимизировать вероятность для заведомо ошибочных переводов. Такие отрицательные примеры формируются искусственно по заранее определённым шаблонам, моделирующим типовые ошибки, например, пропуски слов, перестановки, дублирование или искажения.

В рассмотренной реализации негативный пример y_- формируется из эталонного перевода y удалением случайного слова. Такая операция позволяет искусственно смоделировать ошибку пропуска, которая достаточно часто возникает в машинных переводах на практике, когда модель не может корректно

обработать длинные или сложные конструкции. Формально, генерацию синтетических негативных примеров описывает следующая процедура:

$$t(y) = y_1..y_{t-1}y_{t+1}...y_{|y|}, \quad t \sim \mathcal{U}[1, ..., |y|],$$

где y состоит из слов $y_1...y_{|y|}$, $|y|$ — длина текста y, а t выбирается случайно из целых чисел от 1 до $|y|$.

После построения пары (y, y_-) к каждому исходному предложению x применяется контрастная функция потерь, которая сравнивает логарифмы вероятностей положительного и отрицательного переводов, вычисляемых моделью, и минимизирует этот разрыв. По сути, данная функция реализует обучение с ограниченной максимизацией отступа: если вероятность правильного перевода уже значительно превышает вероятность ошибочного — с учётом заданного отступа α, — обновление весов по данному примеру не производится. Функция потерь может быть записана следующим образом:

$$L_\alpha(x, y, y_-, \theta) = \max(0, \log P_\theta(y_-|x) - \log P_\theta(y|x) + \alpha), \quad y_- = t(y), \qquad (3.1)$$

Такой подход не только повышает стойкость модели к вставкам и пропускам, но и способствует формированию более устойчивого представления о границах грамматически и семантически приемлемых преобразований. По мере усложнения шаблонов для генерации негативных примеров и выбора оптимального значения отступа α можно варьировать чувствительность модели к разным видам ошибок, чем обеспечивается гибкая настройка системы под потребности целевой задачи и специфику используемых параллельных данных.

Однако при дообучении модели только с использованием контрастной функции потерь 3.1 возникает существенный риск «забывания» основной задачи — авторегрессионной генерации перевода. Поскольку контрастная функция потерь нацелена исключительно на различение правильных и ошибочных переводов на уровне предложения, она не контролирует корректность пословных предсказаний $P_\theta(y_t|y_{<t}, x)$, которые лежат в основе итеративной генерации перевода в авторегрессионных моделях. В результате воздействие обучения становится слишком «грубым»: модель может терять способность поэтапно строить перевод, что приводит к ухудшению итогового качества или даже полной утрате способности к последовательной генерации.

Для того чтобы избежать этой проблемы, в данной работе была выбрана комбинированная функция потерь, представляющая собой линейную комбинацию авторегрессионной компоненты, которая отвечает за вероятностную правильность генерации каждого слова по алгоритму максимизации правдоподобия, и контрастной компоненты, которая стимулирует различение правильных и неправильных переводов. Такая функция потерь можно записать следующим образом:

$$L_{\alpha,\beta} = \beta \log P_\theta(y|x) + \max(0, \log P_\theta(y_-|x) - \log P_\theta(y|x) + \alpha). \qquad (3.2)$$

Подобная постановка позволяет сохранить стабильность и обучаемость авторегрессионного механизма, одновременно интегрируя в процесс обучения для модели информацию о типовых ошибках и негативных примерах за счёт контрастного слагаемого. Параметр β определяет баланс между основным процессом максимизации правдоподобия и дополнительным штрафом за ошибки различения. Экспериментальные данные подтверждают, что комбинированный подход обеспечивает более высокое качество перевода и устойчивость обучения по сравнению с использованием одной только контрастной функции потерь.

Подбор параметров α и β необходимо осуществлять экспериментально. Он будет описан позже. Далее будет показано, что дообучение только на контрастную функцию потерь 3.1 действительно приводит к более плохому результату, чем обучение на линейную комбинацию 3.2. Далее обучение с функцией потерь 3.2 будем называть контрастным обучением.

3.2 Выбор лучшего перевода с помощью разметки, выполненной человеком

В отличие от привычных техник, когда негативные примеры формируются по фиксированным шаблонам, в данной работе предлагается подход, при котором негативные переводы формируются естественным образом из самой работы модели. Для реализации этого подхода с одной и той же модели с помощью процедуры разнообразного поиска в ширину [37] генерируется сразу несколько различных вариантов перевода для одного и того же входного текста. Поскольку все альтернативные переводы создаются работающей моделью, среди них

нередко оказываются достаточно правдоподобные гипотезы, действительно отражающие типичные ошибки или неочевидные варианты передачи смысла. С точки зрения обучения это оказывается значительно более сложным и реалистичным негативным материалом, чем синтетические переводы, испорченные с помощью шаблонных ошибок. Полученные переводы отправляются людям для ранжирования по качеству. После этого исходный текст и отранжированные по качеству гипотезы перевода используются для обучения модели перевода. Для этого используется функция потерь 3.2. Подробно смеху метода можно увидеть на рисунке 3.1.

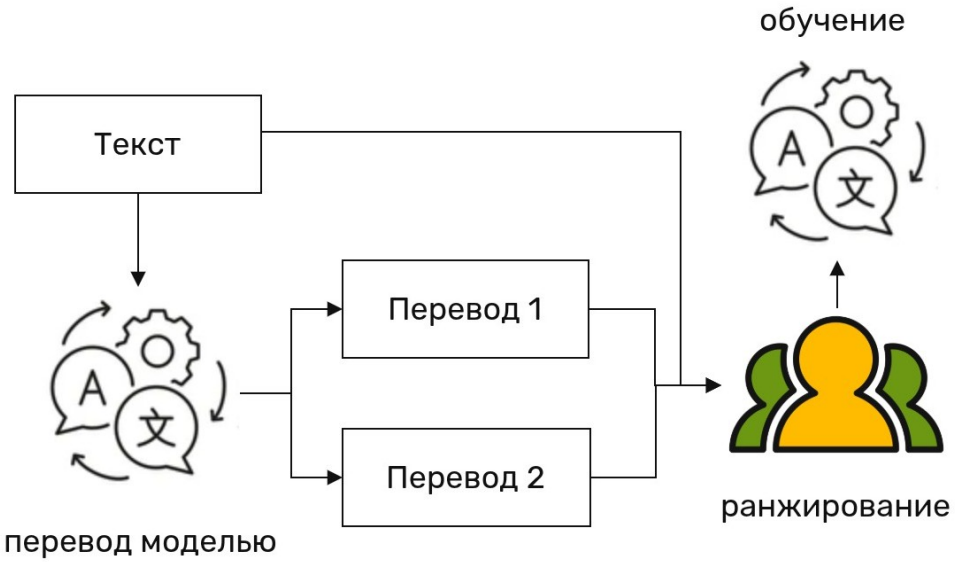


Рисунок 3.1 — Схема метода переупорядочивания гипотез перевода на основе человеческой разметки.

Для корректного сбора обратной связи от экспертов важным этапом становится разработка выверенной инструкции и унификация процедуры разметки. Участникам разметки предлагается для каждой пары сгенерированных переводов сравнить качество переводов по входному тексту и выбрать лучший среди двух либо указать, что переводы равноценны по качеству. Для повышения прозрачности и облегчения процесса оценки различия между двумя переводами автоматически подсвечиваются. Инструкция содержит перечень критериев для выбора лучшего варианта. Пример задания разметки представлен на рисунке 3.2. В инструкции обращается внимание как на точность передачи смысла и терминологию, так и на корректность грамматического оформления, пунктуацию, стилистику и правильную расстановку заглавных букв. При этом список ошибок не подразумевает жёсткой иерархии — размечающим предоставляется свобода

самостоятельно определять, какая ошибка кажется наиболее существенной для качества перевода в каждом конкретном случае. Итогом такой организации процесса становится формирование корпуса упорядоченных по человеческой оценке пар переводов: для каждого исходного текста пара из позитивного и негативного примеров строго согласована с суждением эксперта. Обучение на таком корпусе позволяет делать модель более чувствительной к реальным ошибкам, которые замечает человек, и проводить более точную доменную адаптацию.

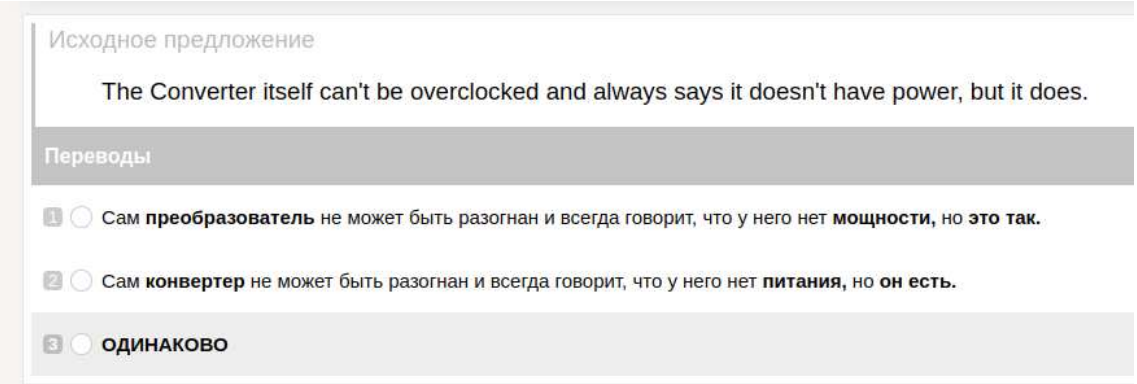


Рисунок 3.2 — Пример задания выбора лучшего перевода для аннотаторов

В целях повышения качества разметки, выполненной человеком, и минимизации субъективных или случайных ошибок в оценках перед началом основной работы был организован обязательный экзамен-допуск для всех участников разметки. Экзамен включал в себя ряд контрольных примеров с заранее подготовленными правильными ответами, основанными на экспертной лингвистической оценке и типовых ошибках перевода. Только успешно справившиеся с этими заданиями для допуска операторы переходили к основной части разметки.

В ходе разметочной сессии для поддержания стабильного уровня качества и предотвращения эффекта усталости или потери концентрации в задании периодически появлялись вкраплённые контрольные вопросы с заранее подготовленными лингвистами ответами. Неверный ответ на такой контрольный вопрос автоматически приводил к приостановке работы размечающего и временной или полной потере права участвовать в дальнейшем разметочном процессе. Такой механизм позволял не только поддерживать высокую достоверность результатов разметки, но и отсекать небрежных или недобросовестных участников, что в итоге существенно улучшило качество итогового корпуса и надёжность эмпирических результатов, полученных на его основе.

Описанный выше процесс разметки позволяет эффективно масштабировать объемы разметки благодаря процедурам контроля качества и экзаменам. Постро-

ение такого процесса разметки может представлять ценность в промышленных системах, где необходимо получать большие объемы оценок качества перевода для специализированной литературы или документации.

3.3 Оценка качества перевода

Для оценки качества переводов дообученных моделей используется автоматическая метрика BLEU [18], описанная в разделе 1.4. Эта метрика требует наличия эталонных переводов для тестового корпуса и показывает достаточно высокую корреляцию с оценками людей.

При проведении экспериментов метрика BLEU вычисляется с эталонными переводами, подготовленными профессиональными переводчиками. Для экспериментов использовались тестовые корпуса для русско-английского направления, подготовленные к конференции WMT-2019 [38]. Размер тестового корпуса составляет 3000 предложений, исходные предложения выбирались из новостных статей.

Кроме автоматической метрики BLEU в данной работе для объективного сравнения обученных моделей перевода используется также оценка на основе разметки, выполненной человеком, процедура которой подробно описана в разделе 3.2. Для минимизации эффекта переобучения модели под специфические предпочтения конкретных людей в оценке качества итоговых моделей участвуют лишь те разметчики, которые не принимали участия в процессе разметки данных для обучения. Такой подход позволяет получить независимую экспертную оценку и обеспечивает более достоверную проверку реального качества перевода.

Разметка, выполненная человеком, особенно важна в ситуациях, когда автоматических эталонов либо не существует, либо они не охватывают всего многообразия допустимых переводов и вариантов передачи смысла. Оценка на основе мнения людей позволяет не только количественно сравнить качество различных моделей, но и оценить те аспекты, которые остаются вне поля зрения формальных метрик, например стилистика, плавность, адекватность передачи сложных конструкций. Несмотря на высокую информативность и степень доверия к результатам ручной оценки, чтобы исключить возможность подстройки под особенности выбранной метрики или разметки, в работе особое внимание уделяется одновременной оценке переводов по BLEU и независимой человеческой

оценке. Такой комплексный подход обеспечивает более всесторонний и объективный анализ эффективности и качества современных систем машинного перевода.

Помимо прямого сравнения качества переводов различных моделей, особое значение в исследовании имеет анализ влияния процедуры дообучения с использованием упорядочивания гипотез по разметке, выполненной человеком, на снижение доли систематических ошибок. Особый интерес вызывает задача борьбы с ошибками недостаточного перевода — случаями, когда отдельные смысловые фрагменты исходного текста оказываются опущенными или утрачивают свою информативность в процессе генерации перевода.

Для получения количественной оценки этого эффекта часть переводов, сгенерированных различными моделями, дополнительно демонстрировалась независимым разметчикам, не участвовавшим в создании обучающей выборки, с вопросом: «Является ли перевод недостаточным?», то есть отсутствуют ли в нем части информации, присутствующие во входном предложении. Такой формат оценки позволил выявить и зафиксировать ошибки, которые автоматически не детектировались обычными метриками схожести с эталоном, а также провести глубинный анализ характера и причин типовых недостатков. Это дало возможность не только сопоставить обычные машинные метрики BLEU с человеческим восприятием качества, но и непосредственно замерить уменьшение доли систематических ошибок после дообучения модели на корпусе, ранжированном экспертами по качеству.

3.4 Эксперименты с переранжированием гипотез на основе разметки, выполненной человеком

В данной части главы подробно рассматривается экспериментальное исследование эффективности методов дообучения моделей машинного перевода с использованием переупорядочивания гипотез по разметке, выполненной человеком. В начале излагается описание используемой в экспериментах архитектуры модели, условий и схемы обучения, включая процесс подбора гиперпараметров для различных функций потерь, что позволяет объективно сравнивать различные подходы по единой методологии. Далее приводится описание исходных данных — как параллельных корпусов, так и специально собранного корпуса переводов

с ранжированием по качеству, выполненным людьми, а также методики формирования обучающих примеров и отбора данных для разметки.

Затем анализируются и сравниваются разные схемы дообучения: как традиционное дообучение с использованием только параллельных данных, так и различные варианты контрастного обучения, как с искусственно сгенерированными негативными примерами, так и с использованием оценки качества перевода, выполненной людьми. Особое внимание уделяется сравнению моделей как по автоматическим метрикам, так и по количественной оценке доли недостаточных переводов, выявленных экспертами, а также анализу конкретных примеров, иллюстрирующих преимущества рассматриваемого подхода.

Завершается часть рассмотрением задачи доменной адаптации. Демонстрируется, как предложенный соискателем метод дообучения с разметкой, выполненной человеком, позволяет гибко и эффективно улучшать качество перевода для специализированных текстов на примере товарных заголовков из домена электронной коммерции даже при отсутствии параллельных эталонных данных. Это делает обсуждаемые методы универсальными и полезными на практике для широкого круга реальных приложений машинного перевода.

3.4.1 Архитектура модели

В ходе экспериментов проводилось дообучение модели машинного перевода с использованием различных функций потерь, а также на основе дополнительных данных, размеченных экспертами. В качестве исходной, предобученной модели была выбрана нейросетевая модель из библиотеки fairseq, которая на момент проведения экспериментов являлась передовой по качеству решения задачи перевода с русского на английский язык согласно результатам соревнования WMT-2019 [38]. Эта модель построена на основе архитектуры Transformer-big [2], описание которой приводится в разделе 1.3.

Архитектура Transformer-big отличается увеличенными размерностями внутренних представлений, что способствует более глубокому захвату контекстных связей в тексте и формированию сложных смысловых зависимостей между словами и фразами. Для представления входных последовательностей модель использует векторное пространство размерности 1024, а размерность внутрен-

них слоёв FFN сетей увеличена до 4096, что позволяет обрабатывать более сложные паттерны и ассоциации в языковых данных. Кроме того, в многоголовочном механизме внимания используется 16 параллельных «голов», что даёт модели возможность учитывать различные аспекты языкового контекста одновременно и повышает качество моделирования зависимостей между различными участками текста.

Процесс дообучения реализовывался с использованием оптимизатора Adam [24], который является одним из наиболее устойчивых и эффективных для глубоких нейронных сетей. Для плавного вхождения модели в режим интенсивного обновления параметров применялась стратегия нагрева: в течение первых 1000 итераций темп обучения постепенно увеличивался, что позволяло сгладить возможные скачки градиентов и избежать неустойчивости на ранних этапах. По достижении максимального значения темп обучения снижался по корню из числа оставшихся шагов в течение оставшихся 4000 итераций. Такой подход к настройке параметров обучения способствовал устойчивой и быстрой сходимости процесса оптимизации.

Все эксперименты выполнялись на современном вычислительном сервере, оснащённом четырьмя графическими ускорителями Tesla M40. Это обеспечивало высокий уровень параллелизма и позволяло за разумное время проводить обучение даже для объёмных моделей с большим числом параметров и на обширных обучающих выборках. Такой вычислительный ресурс, в сочетании с передовой архитектурой модели и гибкими стратегиями оптимизации, созданными на основе современных научных достижений, обеспечил надёжную экспериментальную платформу для всестороннего анализа эффективности различных функций потерь и методов дообучения нейросетевых моделей машинного перевода.

3.4.2 Данные для обучения

Эксперименты, как уже было описано, проводились на направлении с русского на английский язык. Предобученная модель обучалась на данных, предоставленных для соревнования WMT-2019 [38], которые состоят из данных Paracrawl v3, Common Crawl, News Commentary и других корпусов. Для дообу-

чения использовались данные разметки, осуществленной на основе переводов текстов из датасета News Commentary.

Для разметки были выбраны случайно 10000 текстов из указанного обучающего набора. Для каждого из них были составлены два перевода предобученной модели с помощью процедуры разнообразного поиска в ширину [37]. Далее тексты, у которы длина входного текста и перевода не превышает три слова, отфильтровывались. Из оставшихся переводов выбирался лучший на основе процедуры, указанной в разделе 3.2.

Из полученных данных разметки выбирались только те примеры, где размечающие выбрали какой-либо перевод в качестве лучшего. Оказалось, что из всех обучающих примеров в 33% случаев по оценке человека перевод, который был менее вероятен с точки зрения модели, оказывался лучше, чем более вероятный. Это свидетельствует о том, что без дообучения на разметку, выполненную человеком, ранжирования гипотез модели по ней самой не совпадают с ранжированиями размечающих. В 15% случаев с точки зрения размечающих качество оказывалось одинаковым. Эти примеры не использовались для дообучения моделей, так как их не удалось упорядочить по качеству.

3.4.3 Эксперименты с контрастной функцией потерь

Для выбора гиперпараметров отступа α и веса β в контрастной функции потерь 3.2, а также гиперпараметра отступа α в контрастном обучении с сгенерированными по шаблону негативными примерами 3.1 были обучены модели с перебором гиперпараметров по сеткам. При отборе моделей осуществлялся выбор лучшей конфигурации по метрике BLEU на датасете WMT-17 с более старого соревнования по машинному переводу. Отбор моделей по WMT-19 не проводился, чтобы избежать переобучения под тестовую выборку.

Для контрастного обучения 3.2 оптимальные значения гиперпараметров оказались $\alpha = 0.3$, $\beta = 0.1$. Для обучения с негативными примерами, сгенерированными по шаблону 3.1, оптимальное значение α оказалось равным 1.0.

3.5 Эксперименты с моделью переупорядочивания гипотез с помощью разметки, выполненной человеком

Для оценки описанных подходов были обучены следующие модели:

базовая модель, взятая из библиотеки fairseq, которая дообучалась в остальных экспериментах;

дообученная на параллельные данные модель, которая училась столько же шагов, сколько и остальные дообученные модели с авторегрессионной функцией потерь 1.1;

дообученная на победителя разметки модель, которая дообучалась с авторегрессионной функцией потерь 1.1 на тот перевод, который победил в разметке;

дообученная с шаблонными негативными примерами модель, которая дообучалась с контрастной функцией потерь 3.1 с негативными примерами, которые генерировались по шаблону;

дообученная на разметку модель, которая обучалась с контрастной функцией потерь 3.2 на упорядоченные по качеству с помощью разметки, выполненной человеком, переводы.

Результаты оценки обученных моделей по BLEU на тестовом наборе WMT-19 представлен в таблице 4. Как можно заметить, обе модели, которые дообучались на данные, полученные с помощью разметки на качество, выполненной человеком, имеют значительный прирост на тестовом наборе. Причем модель, которая дообучалась с контрастной функцией потерь 3.2, превосходит по BLEU модель, обучавшуюся только на победителя без негативных примеров, а ее прирост относительно базовой модели перевода, равный 1.6 пункта BLEU, максимальный среди всех сравниваемых моделей. Согласно последним исследованиям метрики BLEU такие приросты имеют значимую корреляцию с разметкой качества, выполненной людьми [34].

Сравнение с качеством перевода Google Translate приведено для ориентира. Выводы на основе этого сравнения сделать затруднительно, так как неизвестны характеристики данной системы, а качество системы меняется со временем.

Особое внимание в проведённом исследовании уделялось анализу того, насколько обучение с привлечением разметки, выполненной человеком, способствует устранению систематических ошибок, характерных для нейронных

Таблица 4 — Сравнение дообученных моделей на тестовом наборе WMT19 на направлении с русского на английский по метрике BLEU.

Модель	BLEU
Базовая модель перевода	35.6
Дообучение на переводные данные 1.1	35.7
Дообучение на победителя разметки 1.1	36.7
Контрастное дообучение с шаблонными негативами 3.1	35.8
Контрастное дообучение на разметку 3.2	**37.3**
Google Translate (2025)	40.1

моделей перевода. Основной категорией подобных ошибок выступают так называемые недостаточные переводы — ситуации, когда некоторые значимые фрагменты содержимого исходного предложения утрачиваются в переводном тексте. Для объективной оценки данного аспекта проводился дополнительный этап аннотирования: независимым экспертам предлагалось определить, полностью ли передана информация из оригинального предложения, или, напротив, в переводе обнаруживаются опущенные детали.

Проведённый анализ позволил установить, что дообучение модели на корпусе, ранжированном живыми экспертами, является эффективным способом устранения подобных ошибок. Это можно объяснить тем, что сами по себе негативные примеры, автоматически отобранные из альтернативных гипотез работы модели, зачастую оказываются значительно более сложными и разнообразными, чем примеры, формируемые с помощью формальных шаблонов. В количественном выражении это отражается существенным снижением доли недостаточных переводов: после контрастного дообучения на разметке 3.2 частота таких ошибок снижается с 4% до 1%. Для сравнения, если используются только шаблонно-сгенерированные негативные примеры 3.1, доля недостаточных переводов уменьшается не столь значительно — лишь до 2%. Эти результаты представлены в таблице 5.

Дополнительную информацию о качестве перевода позволила получить независимая оценка самих текстов людьми, проводившаяся по процедуре, подробно описанной в разделе 3.2. При этом для чистоты результата к участию в оценке были допущены только те эксперты, которые не принимали участия в обучающей разметке. В результате такой проверки выяснилось, что дообучение на разметку 3.2 способствует заметному улучшению качества перевода: по

Таблица 5 — Сравнение дообученных моделей на тестовом наборе WMT19 на направлении с русского на английский по доле недостаточных переводов.

Модель	Доля недостаточных переводов, %
Базовая модель перевода	4
Контрастное дообучение с шаблонными негативами 3.1	2
Контрастное дообучение на разметку 3.2	**1**

заключению размечающих, оно даёт улучшение в 15% случаев по сравнению с исходной моделью. Для сравнения, дообучение только с использованием шаблонных негативов даёт положительный эффект лишь в 3% случаев. Примеры переводов, которые были признаны существенно улучшенными после контрастного дообучения на разметке, выполненной человеком, представлены в таблице 6. В таблице приведены несколько примеров, демонстрирующих различные систематические ошибки перевода: потеря гладкости перевода и потеря смысла.

Таблица 6 — Примеры переводов моделей. Входом обозначены тексты, подаваемые на вход моделям. Базовой называется модель до дообучения на разметку 3.2. Дообученной обозначена модель после контрастного дообучения на разметку 3.2, выполненную человеком

Вход:	а в одиночку как-то скучно
Базовая:	and alone as it is boring
Дообученная:	and it’s kind of boring alone

Вход:	все время загорается красным
Базовая:	it’s always red
Дообученная:	it lights up red all the time

Вход:	у меня немного опыта путешествий
Базовая:	I don’t have much experience
Дообученная:	I have a little experience in traveling

3.5.1 Эксперименты с доменной адаптацией

Рассмотрим подробнее вопрос о применимости предлагаемого подхода с контрастным обучением на основе разметки 3.2 для задач доменной адаптации нейронных моделей машинного перевода. Как уже отмечалось ранее, одно из важных преимуществ использования разметки, выполненной человеком, заключается в том, что данный подход не требует наличия больших и качественных параллельных корпусов в целевом домене. Благодаря этому открывается возможность дообучения моделей под новые специализированные области, для которых подготовка эталонных данных затруднена или экономически нецелесообразна.

Для проверки эффективности этого подхода в реальных прикладных условиях был выбран специфический сценарий — перевод с английского на русский язык в домене электронных коммерческих платформ, а именно перевод кратких заголовков товаров из интернет-магазинов. Заголовки товаров — чрезвычайно сложный для машинного перевода тип текстов. Они обладают ярко выраженной спецификой: как правило, лишены стандартной синтаксической структуры, часто представляют собой цепочку ключевых слов, определений и перечислений, и могут включать термины, бренды, технические характеристики. Кроме того, качественные параллельные корпуса для этого домена, как правило, отсутствуют в открытом доступе либо представлены в небольших объёмах.

В рамках эксперимента в качестве исходной использовалась предобученная модель, прошедшая обучение по стандартной авторегрессионной функции потерь 1.1 на больших корпусах общего назначения. Для адаптации к домену заголовков товаров был проведён отбор 10 000 уникальных наименований, к каждому из которых предобученной моделью было сгенерировано несколько вариантов перевода с использованием процедур разнообразного поиска. Далее, как описано в разделе 3.2, для каждой пары альтернативных переводов осуществлялась разметка экспертами с указанием лучшего варианта.

В результате получился специализированный корпус пар переводов, отранжированных по качеству, который использовался для дообучения переводческой модели методом контрастного обучения на разметке, выполненной человеком. Такой подход позволил быстро адаптировать модель к требованиям домена без использования параллельных данных и обеспечить видимый прирост качества по оценке людей. В частности, на примерах из таблицы 7 видно, что исходная модель

зачастую строит переводы с ошибочными структурами, лишними или пропущенными деталями, особенно при отсутствии явного сказуемого или при большом количестве дополняющих определений. После дообучения текст становится значительно более гладким, информативным и соответствующим принятым в домене паттернам описания товаров.

Таблица 7 — Пример переводов моделей при доменной адаптации. Входом обозначены тексты, подаваемые на вход моделям. Базовой называется модель до дообучения на разметку 3.2. Дообученной обозначена модель после контрастного дообучения на разметку 3.2, выполненную человеком.

Вход:	car temporary parking card luminous calling phone number cards with sucker plate
Базовая:	автомобильная временная парковочная карта, светящиеся карточки с номером телефона для звонков с присоской
Дообученная:	светящиеся карточки с номерами телефонов для временной парковки автомобилей с присоской

Так как для данного домена отсутствуют качественные тестовые наборы, оценка качества проводилась с помощью разметки переводов, выполненной людьми. Для этого было взято 1000 заголовков товаров, они были переведены моделями до и после применения предложенного соискателем метода. По итогам разметки, выполненной людьми, оказалось, что качество после дообучения методом переранжирования гипотез 3.2 выросло на 40% заголовков. В табл. 7 представлены примеры того, как улучшился перевод заголовков после процедуры дообучения. В среднем после дообучения переводы стали более гладкими с точки зрения русского языка, и части заголовков стали переводиться согласованно друг с другом.

3.6 Переупорядочивание гипотез в задаче видеоперевода с использованием языковых моделей

В данном разделе представлено описание системы автоматического перевода субтитров, представленную на конференции WMT24 [59; 39]. В исследовании выбрано направление перевода с английского на русский. Подход заключается

в адаптации одной из моделей семейства YandexGPT для решения задач перевода с использованием многоэтапного процесса, обеспечивающего высокое качество и контекстуальную точность перевода. Одним из ключевых внедрений, позволяющих получить высокую точность и гладкость перевода, является применение подхода с переранжированием гипотез на основе разметки [54] к большим языковым моделям (БЯМ), обучаемым на задачу перевода с контекстом. Вклад диссертанта в данную работу заключается в проведении экспериментов с адаптацией переупорядочивания гипотез на основе разметки корпуса предложений.

3.6.1 Описание модели перевода

Использование больших языковых моделей в задаче машинного перевода становится всё более оправданным и естественным с точки зрения как практических результатов, так и фундаментальных свойств самих моделей. Во-первых, языковые модели обучены на колоссальных массивах разнородных текстовых данных на множестве языков, что позволяет им формировать глубоко структурированные представления о грамматике, лексике, семантике и стилистике разных языков, а также закономерностях их взаимодействия. Это важнейшее преимущество по сравнению с классическими моделями машинного перевода, ориентированными только на параллельные корпуса, которые теряют эффективность при недостатке данных или в новых доменах.

Во-вторых, современная языковая модель — это универсальный генератор текстов, способный не только имитировать правила конкретного языка, но и связывать сложные контекстные зависимости на уровне дискурса, учитывать межфразовые и междокументные нюансы, а также обеспечивать когерентность и связность длинных переводимых документов. Благодаря обучению на данных разных языков, такие модели изначально обладают знанием о сопоставимости смыслов между языками, что способствует улучшению эквивалентности переводов даже для языковых пар с недостатком параллельных данных.

Кроме того, языковые модели легко адаптируются под различные форматы задач с помощью техники инструкций, что расширяет их возможности в машинном переводе: перевод отдельных фраз, абзацев, диалогов, строк кода, HTML, специализированных терминов, имен собственных. Возможности интеграции

дополнительных инструкций или уточнений позволяют реализовать кастомизированные подходы к переводу, ориентированные на специфику домена, стиль или корпоративные стандарты, не требуя создания отдельной специализированной архитектуры для каждой разновидности задачи.

Наконец, архитектура языковых моделей изначально строится на механизмах внимания и трансформерных блоках, что обеспечивает высокий уровень параллелизма, масштабируемость и возможность эффективной работы с большими входными текстами. Это не только ускоряет обучение и вывод, но и позволяет расширять границы применения машинного перевода — переходить к задачам многоязычного поиска, суммаризации, постредактирования, генерации парафраз и кросс-лингвистического анализа. Все эти свойства делают использование языковых моделей оправданным, перспективным и практически выгодным решением для задач современного машинного перевода в промышленности и науке.

С другой стороны, современные языковые модели не всегда превосходят специализированные модели перевода [39], так как являются моделями общего назначения. Для более качественного перевода на многих направлениях всё ещё нужна тонкая донастройка модели исключительно под задачу перевода.

В основе предложенного соискателем подхода находится большая языковая модель, обученная на разных языках, с заметной долей русскоязычных и англоязычных данных в обучении. Качество этой предварительно подготовленной модели оценивается с использованием широкого спектра критериев, включая как автоматизированные метрики, так и оценку людьми. Наличие этого начального этапа гарантирует, что модель усвоит широкий спектр лингвистических особенностей и нюансов разных языков, тем самым создав прочную основу для последующей тонкой настройки под задачу машинного перевода.

После этапа предварительного обучения модель готовится к задаче перевода, с помощью добавления параллельных данных, в которых тексты на английском и русском языках объединены с помощью разделителя. Этот шаг крайне важен для того, чтобы модель одинаково хорошо понимала оба языка в контексте перевода. В обучении используется закрытый набор параллельных данных, полученный из веб-документов, аналогичный CommonCrawl [40].

Данные тщательно обработаны для обеспечения высокого качества данные обрабатываются, по процессу, подобному Bicleaner [31], включающему следующие этапы:

– Тексты отбираются с использованием автоматических фильтров параллелизма.
– Дубликаты удаляются для поддержания чистоты набора данных.

Такая стратегия построения корпуса позволяет языковой модели устанавливать связи между двумя языками и изучать прямые соответствия между английским и русским языками и наоборот. После обучения на таких данных, языковая модель можно рассматривать как модель перевода в авторегрессионной постановке задачи 1.1.

Последующим этапом обучения модели было переупорядочивание гипотез на основе разметки, выполненной человеком, с контрастной функцией потерь 3.2. Для этой процедуры метод был адаптирован для применения к обучению языковых моделей. Подробное описание экспериментов с адаптацией метода будет представлено в следующем разделе.

Следующими этапами подготовки модели служили эксперименты с функцией потерь, адаптация метода обучения к контекстному переводе параграфов и адаптация модели к структурированному переводу текстов с тэгами субтитров. Эти этапы конвейера, в том числе обработка структуры абзацев, интеграция и поддержка тегов HTML при генерации, а также корректировка хронометража и временных границ для субтитров и медиаконтента, остаются важными для построения общего пайплайна, однако не рассматриваются подробно в настоящей работе, поскольку соответствующие экспериментальные исследования были проведены вне основной сферы задач диссертации.

3.6.2 Адаптация к языковым моделям метода переупорядочивания гипотез на основе разметки, выполненной человеком

.

Рассмотрим подробнее этап адаптации метода переупорядочивания гипотез модели на основе разметки. В качестве базовой модели для сравнения используется подготовленная к задаче перевода языковую модель, которая прошла две стадии: стадию предварительного обучения с повышенной долей переводных данных и стадию дообучения на отобранных качественных параллельных дан-

ных. Аналогично моделям перевода, обученным методом 1.1, такая модель уже является моделью перевода.

Этап дообучения на отобранные переводные данные проводится на небольшом количестве данных высокого качества в режиме SFT. Данные представляют собой параллельные фрагменты книг длиной до 1000 токенов. Дообучение проводится с использованием техники p-tune [41] — добавлением небольших адаптеров, которые подаются в контекст модели. В процессе дообучения веса сети остаются замороженными и изменения происходят только в адаптерах. Размер каждого блока p-tune равен 100. В целом, входные данные для языковой модели состоят из исходного текста на английском языке, окружённого двумя блоками-адаптерами. Обучение адаптеров позволяет модели избежать переобучения при обучении на небольшом количестве высококачественных данных.

Так как размер и конфигурация языковой модели существенно отличается, для стадии дообучения на ранжированиях, выполненные людьми, необходим тщательный подбор гиперпараметров. Так, для языковой модели заново подбираются: коэффициент скорости обучения, длина обучения, значение гиперпараметра α в функции потерь 3.2, значение ширины при выполнении поиска в ширину при генерации переводов.

Для всесторонней оценки качества перевода и комплексного анализа вклада каждого этапа конвейера принимается решение использовать не только традиционную метрику BLEU, но и современные нейронные метрики, такие как BLEURT-20 [19] и COMET [20]. Такой подход обусловлен тем, что автоматические метрики на основе n-грамм, такие как BLEU, ROUGE, несмотря на их популярность и простоту вычисления, лишь частично коррелируют с субъективными оценками профессиональных переводчиков и часто оказываются недостаточно чувствительными к тонким ошибкам передачи смысла, стилистике или характерным для нейросетевых моделей лингвистическим искажениями. Особенно эти недостатки становятся заметны при работе с системами высокого качества. В связи с этим современные исследования, например [42], рекомендуют опираться, прежде всего, на нейронные метрики, которые демонстрируют более высокую корреляцию с человеческой экспертизой и способны улавливать сложные парафразные или контекстуальные несоответствия.

Метрика BLEURT-20 представляет собой модель, основанную на дообученном на оценках, выполненных человеком, BERT-подобном трансформере, которая по входным данным в виде пары сгенерированный перевод и эталонный перевод

предсказывает скалярный балл качества перевода. BLEURT учится различать не только точные совпадения отдельных слов или фраз, но и учитывать смысловые соответствия, длинные зависимости, грамматические и стилистические особенности текста. Благодаря этому метрика демонстрирует высокую согласованность с экспертными суждениями даже при сравнении переводов, сильно отличающихся по формулировке от эталона.

COMET — другая нейронная метрика, специально разрабатываемая и дообучающаяся для задач машинного перевода, также показывает выдающиеся результаты в задаче оценки качества. Она использует архитектуру мульти-языкового трансформера и обучается на больших выборках реальных оценок переводов, выполненных людьми. В отличие от BLEURT, COMET может работать не только с парой перевод и эталон, но и с исходным предложением на языке-источнике, что позволяет даже лучше учитывать адекватность передачи смысла и контекстуальные соответствия между языками.

Таким образом, применение нейронных метрик позволяет получить более надёжную и точную оценку различий между сравниваемыми моделями и этапами обучения, а их использование в комплексе с классическими метриками обеспечивает комплексный, многоуровневый анализ качества современных систем машинного перевода.

Выбор гиперпараметров дообучения. Обучающий корпус размеченных ранжирований переводов предложений составляет 100000 триплетов. Для подбора гиперпараметров метода переупорядочивания гипотез на основе разметки осуществляется перебор адекватных значений по ограниченной сетке. Количество проверяемых конфигураций можно менять в зависимости от доступных для экспериментов вычислительных ресурсов. Подбор гиперпараметра α и способа выбора итерации обучения представлены в таблице 9. Все эксперименты в этой серии проводились с длиной обучения 2000 итераций.

На основе этих результатов видно, что оптимальным является выбор последней итерации обучения, а оптимальное значение α составляет 5.5. Однако итоговое качество перевода с точки зрения обеих метрик на тестовом наборе WMT21 продолжает расти по мере увеличения значения гиперпараметра α.

В следующей серии экспериментов осуществляется подбор длины обучения. Модель проявляет высокую чувствительность к изменению коэффициента скорости обучения и нечувствительна к изменению длины обучения в исследуемом диапазоне.

Таблица 8 — Подбор значения α при настройке метода переупорядочивания гипотез на основе разметки для языковой модели. Оценка проводится на тестовом наборе WMT21.

α, выбранная итерация модели	BLEURT	COMET
$\alpha = 1.0$, лучшая	73.9	84.3
$\alpha = 1.0$, последняя	74.4	84.5
$\alpha = 3.5$, последняя	74.2	84.6
$\alpha = 5.5$, лучшая	74.2	84.5
$\alpha = 5.5$, последняя	**74.6**	**84.7**

Таблица 9 — Подбор значений длины и коэффициента скорости обучения при настройке метода переупорядочивания гипотез на основе разметки для языковой модели. Оценка проводится на тестовом наборе WMT21.

Длина обучения	Коэф. скорости обучения	BLEURT	COMET
500 шагов	5e-6	74.0	84.4
1000 шагов	5e-6	**74.7**	84.6
2000 шагов	5e-6	74.4	84.4
5000 шагов	1e-6	67.6	77.0
5000 шагов	5e-6	74.0	**84.5**

В качестве оптимальных значений длины и коэффициента скорости обучения выбраны 2000 и 5e-6 соответственно.

Далее осуществляется выбор оптимального значения ширины поиска при генерации. Предполагается, что оптимальное значение может быть больше для языковых моделей, так как они способны генерировать более вариативные тексты благодаря более глубоким представлениям о структуре языка. Результаты исследования представлены в таблице 10.

Таблица 10 — Подбор значения ширины поиска при настройке метода переупорядочивания гипотез на основе разметки для языковой модели. Оценка проводится на тестовом наборе WMT21.

Ширина поиска в режиме генерации	BLEURT	COMET
4	75.3	85.0
8	75.3	85.1
12	**75.4**	**85.1**

Как можно заметить, ширина поиска при генерации не влияет заметно на качество перевода. Однако в отличие от специализированных моделей перевода, качество перевода языковой модели не деградирует по мере увеличения ширины поиска.

Во всех экспериментах выше значения гиперпараметра β из 3.2 оставалось равным 0.1.

Исследование эффективности метода переупорядочивания гипотез. После исследования влияния гиперпараметров остаётся вопрос об эффективности применения метода переупорядочивания гипотез на основе разметки к языковым моделям. Сравнение модели, обученной данным методом 3.2, с моделью, обученной только на отобранном корпусе переводов методом 1.1, представлено в таблице 11. Согласно последним исследованиям метрик COMET и BLEURT такой прирост данных метрик значимо скоррелирован с победой на разметке на качество, выполненной человеком [34].

Таблица 11 — Результаты обучения языковой модели с и без переупорядочивания гипотез на основе разметки, выполненной человеком.

Модель / Этап	BLEURT	COMET
Языковая модель без дообучения 3.2	73.2	83.1
Языковая модель с дообучением 3.2	**75.4**	**85.1**

Итоговое качество модели с названием «Yandex», одним из этапов обучения которой была стадия обучения предложенным соискателем методом переупорядочивания гипотез на основе разметки, выполненной человеком, представлено в результатах соревнования WMT24 [**findings**]. В таблице результатов указан диапазон мест, внутри которого системы перевода не отличимы по качеству с точки зрения человека. 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, среди которых есть Llama3 70B.

Дальнейшие стадии. В рамках подготовки решения для соревнования WMT модель была доработана с помощью нескольких этапов. Так как эксперименты с дальнейшими стадиями лежат вне поля исследования данной диссертации, результаты и суть последующих шагов в данной работе изложены не будут. С более глубокими деталями исследования можно ознакомиться в работе [59].

3.7 Выводы

По итогам экспериментов с методом переупорядочивания гипотез перевода на основе разметки, выполненной человеком, можно утверждать, что предложенный соискателем подход:

- дает прирост по метрике BLEU на корпусе WMT-19 в задаче перевода с русского на английский по сравнению с базовым авторегрессионным обучением. Качественный прирост подтверждается разметкой, проведенной людьми, на которой видно, что доля систематических ошибок при переводе становится меньше. Также предложенный соискателем подход превосходит по качеству метод обучения с шаблонными негативными примерами [8];
- успешно переносится на задачу перевода специализированных текстов на примере домена электронной коммерции. Прирост подтверждается значимым ростом качества перевода заголовков товаров с точки зрения разметки, выполненной людьми;
- успешно переносится на современные языковые модели, адаптированные к задаче перевода, и дает улучшение по автоматическим метрикам качества перевода, таким как BLEURT и COMET.

Глава 4. Разработка метода маскировки входа для улучшения качества перевода

В современном мире технологии машинного перевода играют ключевую роль в облегчении коммуникации между людьми, говорящими на разных языках. Однако несмотря на значительные успехи в этой области, задача перевода до сих пор не решена.

Так, переводы онлайн-систем и академических автоматических переводчиков обладают некоторым набором свойственных им проблем, например неточность перевода или его неграмотность [8]. Источниками этих ошибок могут являться как некачественно подготовленные данные, так и несовершенство выбранной функции потерь, которая может недостаточно сильно штрафовать модель за неточности в переводе.

Есть гипотеза, что систематические проблемы берут истоки в некачественных переводных данных, на которых учится модель [54]. Плохо выровненные переводы, полученные автоматическими методами, смещают модель. В итоге в некоторых генерациях моделей наблюдаются избыточные и недостаточные переводы. Для борьбы с этими ошибками предлагается метод переупорядочивания гипотез перевода на основе разметки, выполненной человеком.

В данном исследовании предлагается несколько иной взгляд на проблемы перевода. Так, задача перевода рассматривается как частный случай языкового моделирования. Для языковых моделей известно, что переход от задачи предсказание следующего слова к маскированным функциям потерь дает качественные улучшения. Такие функции потерь стимулируют модель выучивать более сложные контекстные зависимости и такие модели показывают более высокое качество в прикладных задачах анализа текстов, таких как суммаризация и моделирование ответов на вопросы [10].

В данной главе представлен новый метод, основанный на методе маскировки входа для улучшения качества машинного перевода. Предложенный соискателем подход вдохновлен успехами метода маскировки в смежной задаче языкового моделирования и предлагает новую функцию потерь, которая поощряет модель к выучиванию более сложных языковых закономерностей, и, как следствие, приводит к более высокому качеству перевода. Основное предположение формулируется следующим образом: переход к маскированной функции

потерь в обучении приводит к количественному снижению систематических ошибок в переводе.

Для демонстрации эффективности предложенного метода в работе представлены эксперименты на направлении перевода с английского на русский язык. Для оценки качества переводов используются как стандартные для задачи метрики качества, так и новые, способные оценивать лингвистическую приемлемость полученных после перевода текстов. Результаты показывают, что использование метода маскировки входа может улучшить качество модели, а также открывает путь к построению качественных моделей в таких смежных задачах как постредактирование перевода.

Исследования метода маскировки входа для улучшения качества машинного перевода и применение полученных моделей в смежных задачах подробно изложены в работе [56].

4.1 Постановка задачи маскированного языкового моделирования

Языковые модели широко применяются в задачах как генерации, так и анализа текстов на естественном языке, так как способны захватывать сложные языковые закономерности, определять контекстуальные и семантические связи между словами. Это позволяет подобным моделям создавать тексты, которые выглядят естественно и адекватно соответствуют исходному контексту. Однако исследования продемонстрировали, что стандартная задача предсказания следующего слова может быть усложнена введением дополнительных механизмов. Одним из таких механизмов является частичная маскировка символов во входном тексте, записываемая как $m(x, p)$ — здесь на позиции p в тексте x устанавливаются маски:

$$m(x,p) = \alpha^1, \ldots, \alpha^{|x|},$$
$$\alpha^i = \begin{cases} x^i, i \notin p \\ [\text{mask}], i \in p \end{cases},$$

где [mask] обозначает специальный маркер маски. В процессе обучения позиции p для каждого текста выбираются случайным образом, при этом маскиру-

ется фиксированный процент слов. Например, в модели BERT реализована задача восстановления исходных слов по частично замаскированному тексту. Функция потерь для её обучения записывается следующим образом:

$$L_\theta(D) = -\sum_{i=1}^{M} \sum_{y_i^j \in y_i \setminus m(y_i)} \log P_\theta(y_i^j | m(y_i)), \tag{4.1}$$

где маскируется примерно 15% исходного текста y.

Особенностью такой постановки является то, что модель, предсказывая отдельное слово, может опираться не только на левый контекст $y_i^{<j}$, как в классических языковых моделях или в машинном переводе 1.1, но и на незамаскированную часть правого контекста, содержащуюся в $m(y_i)$. Экспериментально этот подход доказал свою пользу для задач понимания текстов [9], однако для задач генерации текстов он не подходит, так как модель не является авторегрессионной.

Для решения именно генеративных задач авторами статьи [10] была предложена модель BART: здесь маскированная часть последовательности подаётся в качестве входа кодировщику, а декодировщик восстанавливает исходный полный текст в авторегрессионном режиме. Функция потерь для такого случая имеет вид:

$$L_\theta(D) = -\sum_{i=1}^{M} \sum_{j=1}^{|y_i|} \log P_\theta(y_i^j | y_i^{<j}, m(y_i)), \tag{4.2}$$

где $y_i^{<j}$ — последовательность токенов до j-го, а $|y|$ — длина текста. Важно, что маскированный правый контекст остаётся доступным в качестве подсказки, а сама задача усложняется по сравнению с обычной языковой моделью.

Ключевая особенность данного подхода в том, что модель способна, опираясь на сложные маскировки, восстанавливать целые фрагменты текста авторегрессионно — и тем самым подходит для генеративных задач, например, суммаризации.

4.2 Метод маскировки входа в моделях машинного перевода

Метод маскировки входа в моделях машинного перевода основан на развитии идей маскированного языкового моделирования с учётом специфики пе-

ревода между двумя языками. В классической формулировке, описанной выше, функция потерь 4.2 обучает модель восстанавливать маскированную часть целевого текста на основе его незамаскированной части. Однако для задачи перевода этого недостаточно, так как здесь критично важно учитывать взаимосвязь между исходным и целевым языком — то есть необходимо, чтобы модель умела строить перевод на целевом языке, опираясь на содержание текста источника.

Для этого в предлагаемом варианте вводится функция потерь, учитывающая зависимость целевого текста не только от его собственного частичного представления, но и от частично замаскированного исходного текста x_i:

$$L_\theta(D) = -\sum_{i=1}^{M}\sum_{j=1}^{|y_i|}\log P_\theta(y_i^j | y_i^{<j}, m_1(y_i), m_2(x_i)), \tag{4.3}$$

Здесь $m_1(y_i)$ и $m_2(x_i)$ — варианты маскировки целевого и исходного текстов соответственно. На этапе обучения кодировщик получает на вход обе последовательности, в которых случайным образом замаскирована значительная доля токенов, вплоть до 50%. Декодировщик, аналогично архитектуре BART, восстанавливает полный перевод целиком в авторегрессионном режиме.

В отличие от классической функции потерь 1.1, в которой модель учится строить последовательный перевод, опираясь только на предшествующие слова и полный исходный текст, метод маскировки входа существенно усложняет задачу. Модель вынуждена «разгадывать» перевод на целевом языке, восстанавливая смысл и структуру текста с опорой на неполную, замаскированную версию исходного предложения. Такой механизм способствует формированию более глубоких контекстных и семантических связей внутри целевого языка, повышая гладкость и естественность результатов перевода. Ещё одним преимуществом подхода является возможность подавать на вход декодировщику маскированный целевой текст, что позволяет учитывать частичный правый контекст и тем самым снижать типичные ошибки авторегрессионных моделей, таких как появление избыточных или недостаточных переводов.

Маскирование токенов, как и в BERT-подобных моделях, происходит случайно по выбранным позициям, а при последовательных масках они агрегируются в одну, чтобы не сообщать модели дополнительную информацию о точном количестве пропущенных элементов. Это изменение ещё более усложняет задачу и стимулирует модель к выучиванию глубоких смысловых закономерностей

внутри текстов. Высокая степень маскировки обеспечивает дополнительное разнообразие обучающих примеров, делая получаемую модель более устойчивой и универсальной для последующих задач машинного перевода.

4.2.1 Дообучение на задачу перевода

Полученная функция потерь 4.3 позволяет подготовить модель, которая по частично замаскированным входному тексту и переводу позволяет построить полный перевод. Для перевода такая модель неприменима, так как при переводе не предоставляются части переведенного текста. Для того, чтобы модель стала переводной после обучения с функцией потерь 4.3, модель дообучается на переводной корпус с функцией потерь 1.1. Такое дообучение является частным случаем функции потерь 4.3, при $m_1(y) \equiv m_0$, а $m_2(x) \equiv x$, где m_0 — последовательность, состоящая только из символа маски. После такого короткого дообучения модель становится способна переводить, при этом большую часть обучения она учится с предположительно более сложной функцией потерь.

После дообучения на задачу перевода для улучшения качества модели применим метод переупорядочивания гипотез на основе разметки [54], выполненной человеком. Используемый метод также помогает бороться с систематическими проблемами в переводе, поэтому потенциал метода маскировки входа целесообразно исследовать совместно с методом переупорядочивания гипотез. В рамках этого подхода все переводные модели в экспериментах будут дообучены на триплеты с положительными и отрицательными примерами перевода с помощью ранжирующей функции потерь:

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

где константы α, β подбираются экспериментально, а примеры y_+ и y_- определяются с помощью разметки на людях.

4.2.2 Дообучение на задачу постредактирования перевода

Предложенная соискателем функция потерь 4.3 обладает универсальным характером и применима не только к задаче собственно машинного перевода. После этапа обучения модель становится способной воспринимать на входе текст как на исходном, так и на целевом языке, а также работать с их сочетаниями. Такой универсализм позволяет использовать одну и ту же архитектуру не только для классического перевода, но и для смежных задач обработки текста, в частности — для задачи постредактирования.

Задачу постредактирования рассмотрим как частный случай общей задачи, решаемой моделью, обученной с функцией потерь 4.3. В рамках постредактирования имеется множество пар (x_i, y_i), где x_i — исходный текст на исходном языке, а y_i — автоматически сгенерированный или предварительно переведённый текст на целевом языке. Для каждого примера требуется не просто повторить этот перевод, а улучшить его, то есть получить скорректированный выходной текст z_i, который будет отображать высокое качество, соответствовать стилистическим и смысловым требованиям, устранять ошибки из чернового перевода.

Для такой задачи оптимальная функция потерь будет выглядеть следующим образом:

$$L(X, Y, Z) = -\sum_{i=1}^{N}\sum_{j=1}^{|z_i|}\log P_\theta(z_i^j | z_i^{<j}, x_i, y_i), \tag{4.4}$$

Здесь выход z_i создаётся моделью на основе сразу двух входов: исходного текста x_i и чернового перевода y_i. Благодаря конструкции функции потерь 4.3, где модель изначально учится работать с частично маскированным целевым текстом, архитектура легко адаптируется к ситуации, когда на вход декодеру подаются исходный и предварительный переводы без дополнительных изменений в структуре сети или алгоритме обучения. Последнее будет осуществляться методом дообучения модели, обученной с функцией потерь 4.3, на функцию потерь 4.4.

Таким образом, функция 4.3 не только расширяет класс решаемых задач и упрощает совместное решение перевода и постредактирования в рамках одной системы, но и позволяет строить более гибкие, адаптивные модели, способные работать с разнообразными источниками входных данных для гибкой генерации, улучшения и корректировки текстов разного назначения.

4.2.3 Использование одноязычных данных

Кроме того, важной особенностью рассматриваемой функции потерь 4.3 является то, что она обобщает подход маскированного языкового моделирования, позволяя рассматривать функцию потерь 4.2 как её частный случай. Действительно, если в выражении 4.3 использовать замаскированный исходный текст $m_2(x) \equiv m_0$, то есть полностью маскировать входной текст или вовсе его исключать, то задача обучения сводится к стандартному маскированному языковому моделированию без учета влияния исходного языка. Благодаря этому становится возможным строить единый обучающий процесс, в котором модель одновременно осваивает задачу перевода на параллельных данных и языковое моделирование на одноязычных текстах.

Такой подход дает значительные преимущества с практической точки зрения. Во-первых, одноязычные корпуса обычно гораздо многочисленнее и разнообразнее, чем параллельные, что позволяет существенно расширить языковое покрытие и разнообразие лингвистических конструкций, с которыми сталкивается модель в процессе обучения. Это ведет к формированию более глубоких и универсальных представлений о структуре, грамматике и семантике целевого языка, что положительно сказывается на качестве перевода, особенно в условиях нехватки параллельных текстов.

Во-вторых, совместное обучение на одноязычных и параллельных данных открывает новые возможности для построения мультитасковых или многоязычных универсальных моделей. В рамках такой архитектуры одна и та же модель способна выполнять задачи как генерации на одном языке, то есть условное языковое моделирование, так и непосредственно переводить тексты между языками. Это делает модель чрезвычайно гибкой и многофункциональной: она становится применимой для широкого спектра задач, включая автоматический перевод, суммаризацию, генерацию парафраз, корректуру, а также предварительное языковое моделирование для различных NLP-приложений.

В данной работе основное внимание уделяется исследованию качества такой модели именно в контексте машинного перевода. Тем не менее, этот подход открывает перспективы для дальнейшего развития и применения в других смежных задачах обработки текстовой информации на разных языках.

4.3 Эксперименты с моделью маскированного перевода

Архитектура. Для проведения экспериментов использовалась модель, построенная на архитектуре Transformer-big. В данной конфигурации размерность внутренних векторных представлений составляла 1024, размерность слоёв внутри FFN-сетей — 4096, а число голов в механизме многоголового внимания — 16.

Обучение модели осуществлялось с использованием оптимизатора Adam [24]. Процесс обучения включал фазу постепенного увеличения шага обучения в течение первых 10000 итераций, после чего шаг обучения уменьшался по правилу sqrt-расписания на протяжении оставшейся части обучения. Все эксперименты проводились на вычислительном сервере с восемью GPU NVIDIA A100, что позволило эффективно обучать модели с большим количеством параметров.

Описанная архитектура использовалась как для обучения базовой модели с функцией потерь 1.1, так и для построения новых моделей, обучавшихся с функцией потерь 4.3 и впоследствии дообучавшихся с использованием функций потерь 1.1 и 4.4.

Данные для обучения. Эксперименты проводятся на направлении перевода с английского на русский. Использовались данные, предоставленные для соревнования WMT-2019 [38], которые включают данные Paracrawl v3, Common Crawl, News Commentary и другие корпуса. Во всех экспериментах данные для обучения использовались одни и те же. Менялась только функция потерь. Для обучения на одноязычных данных использовались русскоязычные тексты из корпуса mC4. Для дообучения моделей методом переупорядочивания гипотез на основе разметки, выполненной человеком, использовался корпус из 10000 триплетов [54]. В качестве исходных текстов использовались новостные предложения из корпуса mC4. Примеры переводов генерировались с помощью сэмплирования с температурой и далее отправлялись на разметку, где человек выбирал, какой из двух переводов лучше.

Оценка лингвистической приемлемости текстов. В данном исследовании поднимается вопрос о естественности переводов, полученных с помощью систем автоматического перевода. Некачественные переводы моделей зачастую определяются на глаз, так как они нарушают общепринятые нормы языка. Однако до сих пор не появилось универсальной автоматической метрики, которая подходила бы для оценки данного свойства переводов. В данной работе в качестве

решения использована автоматическая метрика, полученная методом дообучения модели RuBERT [43] на корпусе лингвистической приемлемости RuCoLA [44]. ROC AUC полученной метрики лингвистической приемлемости на отложенной части выборки RuCoLa составляет 0.87. Каждому автоматическому переводу такая метрика присваивает значение от 0 до 1. Чем значение выше, тем грамотнее полученный перевод. Для оценки модели, автоматические переводы на тестовом корпусе оцениваются с помощью описанной выше метрики и полученные значения усредняются по всем примерам тестового корпуса. Для финальной проверки достоверности предложенной метрики было проведено сравнение нескольких систем автоматического перевода и эталонов из корпуса WMT19, подготовленных профессиональными переводчиками. Удалось показать, что лингвистическая приемлемость профессиональных переводов по данному классификатору статистически значимо превышает переводы моделей машинного перевода. Для корпуса WMT19 на эталонных переводах значение метрики составило 0.89, тогда как для систем машинного перевода значение метрики не превосходило 0.84.

Описание моделей перевода. В дальнейших экспериментах будут сравнены следующие модели:

- Базовая модель перевода (БМП) — модель, предобученная с нуля с функцией потерь 1.1 исключительно на данных перевода. После первой стадии, эта модель уже способна переводить. Далее, данная модель дообучается методом переупорядочивания гипотез с использованием разметки, выполненной людьми.
- Маскированная модель перевода (ММП) — модель, предобученная с нуля с функцией потерь 4.3 исключительно на данных перевода. После первой стадии, данная модель не способна качественно переводить, так как ожидает на вход маскированную часть перевода. Далее, данная модель дообучается 10000 итераций на те же данные перевода с функцией потерь 1.1. После этого модель дообучается на размеченных парах перевода, размеченных людьми аналогично предыдущей модели.
- Маскированная модель перевода с использованием одноязычных данных (ММП с ОД) — модель, обученная с нуля на смеси корпусов: половину обучающих примеров составляют переводные данные с функцией потерь 4.3, половину — одноязычные данные с функцией потерь 4.2. Далее данная модель дообучается на задаче перевода с функцией потерь 1.1 и на разметку, выполненную человеком, аналогично предыдущей модели.

Эксперименты с моделью маскированного перевода. Результаты сравнения моделей по метрике BLEU приведены в таблице 12. Обучение модели перевода с маскированной функцией потерь (ММП) дало лучший результат по сравнению с БМП как по метрике BLEU, так и по метрике лингвистической приемлемости текстов. По ручной оценке переводов также получилось, что число избыточных и недостаточных переводов упало на 1%, однако для статистически значимого результата было размечено недостаточное количество примеров. Учитывая последние результаты исследования метрики BLEU, данный прирост качества нельзя считать значительным, так как при таких значениях прироста нельзя говорить о высокой корреляции метрики с разметкой на качество, выполненной людьми [34].

Сравнение с качеством перевода Google Translate приведено для ориентира. Выводы на основе этого сравнения сделать затруднительно, так как неизвестны характеристики данной системы, а качество системы меняется со временем.

Таблица 12 — Сравнение моделей перевода, предобученных с различными функциями потерь на корпусе WMT-19

Модель	en-ru-wmt19, BLEU	Лингв. приемлемость
БМП	34	0.82
ММП	**34.35**	**0.83**
ММП с ОД	33.5	0.815
Google Translate (2024)	34.7	-

Модель ММП с ОД не показала в экспериментах лучших результатов. Это можно объяснить тем, что для языковой пары англо-русского перевода присутствует достаточное количество обучающих данных, чтобы модель могла выучить языковые закономерности. Предложенный соискателем подход может дать другие результаты для тех направлений, где количество доступных для обучения переводных данных невелико.

Переводы, полученные с помощью модели ММП показывают более высокие средние значения метрики лингвистической приемлемости. Это свидетельствует о том, что функция потерь 4.3 действительно подталкивает модель к лучшему запоминанию контекстных связей. В то же время модель ММП с ОД не показала более высоких результатов по этой метрике, как ожидалось. Это может свидетельствовать о недостаточной полезности одноязычных данных в обучении

или о их недостаточной сложности и разнообразии. Приведенные выше результаты показывают, что метод маскировки входа оказывается эффективным даже при совместном использовании с методом переупорядочивания гипотез на основе разметки [54]. Это является некоторым подтверждением того, что систематические ошибки перевода объясняются не только некачественными автоматически собранными данными, но и недостатками функции потерь 1.1, обычно используемой при обучении моделей перевода. Чтобы подтвердить это утверждение, рассмотрим качество моделей БМП и ММП без дообучения моделей методом переупорядочивания гипотез на основе разметки, выполненной человеком. В таблице 13 приведены результаты такого обучения. Значения метрики BLEU и лингвистической приемлемости текстов ожидаемо уменьшились по сравнению с аналогичными результатами при дообучении с методом переупорядочивания гипотез. Однако прирост метрик у модели ММП стал более явным.

Таблица 13 — Сравнение моделей перевода, предобученных с различными функциями потерь без дообучения методом переупорядочивания гипотез с использованием разметки, выполненной человеком, на корпусе WMT-19

Модель	**en-ru-wmt19, BLEU**	**Лингв. приемлемость**
БМП без МПГ	32.2	0.78
ММП без МПГ	33.1	0.818

На основе полученных результатов видно, что метод маскировки входа и метод переупорядочивания гипотез дают незначительное улучшение качества модели при совместном использовании с точки зрения метрик BLEU и лингвистической приемлемости. Однако в экспериментах удалось показать, что обучение модели перевода с маскированной функцией потерь позволяет получить более высокое итоговое качество перевода при использовании подхода без метода переупорядочивания гипотез. Это подтверждает исходное предположение о влиянии неоптимальной функции потерь при обучении на итоговое качество модели.

Кроме того, предложенный соискателем подход позволяет добавить одноязычные данные в обучение модели перевода, но роста качества таким способом добиться пока не удалось, вероятно, из-за большого количества параллельных данных на англо-русском направлении.

Примеры переводов моделей. Для демонстрации эффективности метода маскирования переводов в таблице 14 приведено несколько примеров переводов БМП, демонстрирующих разные типы систематических ошибок. В первых

двух случаях представлены примеры неточного перевода, где перевод избыточен и недостаточен соответственно. В последнем случае представлен пример лингвистически неприемлемого перевода с точки зрения русского языка. Во всех приведенных случаях модель ММП таких ошибок не совершает.

Таблица 14 — Примеры переводов моделей БМП и ММП

Входной текст	**Перевод**
I don’t have much experience	БМП: у меня немного опыта путешествий
	ММП: у меня немного опыта
Return True if the string is a decimal string	БМП: Возвращает значение True, если является десятичной строкой
	ММП: Возвращает True, если строка представляет собой десятичную строку
Take Me Higher	БМП: Возьми меня выше
	ММП: Подними меня выше

4.4 Эксперименты с моделью постредактирования

Описание моделей в экспериментах с задачей постредактирования. В экспериментах с постредактированием использовалось несколько моделей:

– *Малая базовая модель перевода (малая БМП).* Для демонстрации способностей модели постредактирования обучена малая базовая модель перевода, которая отличается от БМП тем, что в качестве архитектуры для нее взят Transformer-base вместо Transformer-big. Ожидаемо, качество перевода такой модели будет хуже, но именно ее генерации будут подаваться на вход модели постредактирования. Именно с такой постановкой задачи сталкиваются профессиональные переводчики, когда редактируют тексты, переведенные неизвестными автоматическими системами. Данная модель училась на тех же данных, что и БМП;

– *Модель постредактирования.* Данная модель предобучена с нуля с функцией потерь 4.3 на данных перевода. Первая стадия полностью аналогична первой стадии модели ММП. Далее модель дообучена на размеченных

парах переводов, которые ранее использовались для метода переупорядочивания гипотез. Эти данные представляют собой 10000 троек исходного текста, плохого перевода и улучшенного перевода, что и ожидается для обучения модели постредактирования. Дообучается модель постредактирования с использованием функции потерь 4.4.

Результаты экспериментов в задаче постредактирования. Из таблицы 15 видно, что модель постредактирования успешно поднимает качество переводов малой модели почти до уровня ММП на корпусе WMT-19. При этом для обучения модели постредактирования потребовалось существенно меньшее количество данных, чем если бы пришлось учить модель с нуля с функцией потерь 4.4.

Таблица 15 — Качество модели постредактирования в задаче исправления переводов малой модели на корпусе WMT-19

Модель	en-ru-wmt19, BLEU
Слабая МП	30.3
Постредактирование	**34.1**

Такая модель может быть полезна в работе профессиональных переводчиков не только для улучшения качества перевода. Так, при работе с текстами переводчикам часто необходимо внесение правок в перевод из-за доменной специфики переводимого текста. В этих случаях может быть полезно использовать метод сэмплирования исправлений из модели постредактирования. Все такие исправления, в отличие от модели перевода, будут опираться на исходный текст перевода, не перефразируя его слишком сильно. Такие разумные и разнообразные правки могут быть полезны для упрощения работы редакторов.

4.5 Выводы

В данной работе предложен метод маскировки входа для задачи машинного перевода. Показано, что данный метод является обобщением не только для задачи перевода, но и для задач постредактирования и маскированного языкового моделирования. Несмотря на то, что на этапе предобучения модель учится на задачу

восстановления маски, качество перевода после дообучения вырастает, что говорит о том, что предложенная соискателем функция потерь является более сложной и полезной для модели.

Несмотря на то, что предложенный подход позволяет встроить одноязычные данные в обучение модели перевода, добиться прироста качества этим способом не удалось. Эксперименты с одноязычными данными предлагается продолжить в будущем на направлениях перевода, где доступно меньшее количество переведенных текстов для обучения.

Метод маскировки входа при обучении перевода рекомендуется использовать вместе с другими методами улучшения качества перевода, такими как метод переупорядочивания гипотез с использованием разметки, описанный в разделе 3. В экспериментах удалось показать, что совместное использование этих улучшений приводит к более высокому качеству генераций модели. При этом использование метода маскирования входа без метода переупорядочивания гипотез дает более значимый прирост метрики BLEU [18].

Кроме того, в работе представлена адаптация метода маскировки входа для задачи постредактирования. Полученная модель успешно справляется с исправлением переводов более слабой модели переводов. Такие модели могут представлять особый интерес для редакторов и переводчиков, так как способны генерировать несколько альтернативных улучшений текста.

Глава 5. Разработка метода сегментации на основе тематических моделей для улучшения качества контекстного перевода длинных текстов

Несмотря на заметный прогресс в области машинного перевода в последние годы, качество перевода длинных и тематически неоднородных текстов по-прежнему остаётся актуальной проблемой. Одной из причин этого является то, что большинство современных моделей перевода работают либо на уровне отдельных предложений, либо на коротких, произвольно разбитых отрезках текста, не учитывая при этом тематическую целостность документа. Такой подход может приводить к потере связности, неверной интерпретации многозначных выражений и снижению общего качества перевода. В контексте данной работы длинными текстами будем называть текстовые материалы, превышающие по длине несколько абзацев. Ключевая особенность таких текстов для задачи перевода — необходимость сохранения межфразовой и межсегментной связности. Это подразумевает учёт анафорических ссылок, терминологической согласованности, стилистического единообразия и структурной целостности.

В связи с этим особое значение приобретает задача автоматической сегментации текста на однородные по тематике фрагменты перед переводом. Тематические модели, такие как ARTM [13], позволяют выявлять скрытые темы внутри текста и использовать эту информацию для более осмысленного разбиения документа. Сегментация, основанная на тематическом анализе, обеспечивает тематическое единство каждого выделенного сегмента предложений. Это способствует лучшему сохранению контекста и смысловой целостности при последующем переводе.

Тематическое моделирование — это быстро развивающаяся область статистического анализа текста. Тематическая модель раскрывает скрытую семантическую структуру коллекции текстов и находит сжатое представление каждого документа в виде набора тем. С точки зрения статистики, каждая тема — это набор слов или фраз, которые часто встречаются во многих документах. Тематическая репрезентация документа отражает наиболее важную информацию о его семантике и поэтому полезна для многих задач, включая поиск информации, классификацию, категоризацию, обобщение и сегментацию текстов [13]. Несмотря на множество преимуществ, известно, что тематические модели не справляются с моделированием структуры текста внутри документов. Обычно

все темы, представленные в документе, равномерно распределены по тексту. Это тесно связано с предположением о «мешке слов» при моделировании текстов. Это предположение значительно упрощает теоретический вывод, который позволяет получить итеративное решение, известное как EM-алгоритм. Но во многих задачах, таких как анализ больших документов, внутридокументный поиск или диалоговые системы, важно моделировать внутридокументное тематическое поведение с хорошей детализацией. Одной из самых популярных тематических моделей является латентное размещение Дирихле (LDA), предложенное в [14]. LDA — это двухуровневая байесовская генеративная модель, которая предполагает, что распределение тем по словам и распределение документов по темам генерируются на основе априорных распределений Дирихле.

Многие авторы успешно пытались изменить генеративную модель LDA таким образом, чтобы в неё были включены некоторые допущения о структуре текста. Например, в [45] была создана модель под названием senLDA. В этой модели все слова в предложении могут иметь только одну и ту же тематическую метку. В ходе экспериментов эта модель сходилась быстрее, чем LDA, а представление документов, обеспечиваемое этой моделью, успешно дополняло представление LDA при решении задачи классификации документов. В [16] была предложена более сложная модель тематической сегментации на основе LDA (TSM). Она предполагает, что документы состоят из сегментов, темы которых также присутствуют в темах документов. Для моделирования тем сегментов используется процесс Питмана — Йора [46]. Он представляет каждый сегмент в виде китайского ресторана, где клиенты — это слова, блюда — это темы, а столики — это монотематические подмножества слов. Для всех этих моделей любые предположения о структуре текста изменяют генеративную модель, что затрудняет разработку и вывод новых модификаций.

В этой главе предлагается метод, основанный на аддитивной регуляризации тематических моделей (ARTM) [13]. Данный подход позволяет восстановить сегментарную структуру текста в тематической модели. Предполагаемые границы сегментов используются для сокращения количества тем в сегменте. Это делается исходя из предположения, что слова в текстовом фрагменте относятся к одному и тому же небольшому набору тем. В результате тематическая модель разбивает темы на сегменты и повышает разреженность модели. Эту тематическую структуру можно использовать для автоматического внутридокументального анализа. Наконец, предложенные усовершенствования не усложняют структуру обучения

модели и теоретических выводов, но повышают качество, разреженность и интерпретируемость тематической модели. Всё это открывает новые возможности для применения тематических моделей. Для оценки качества сегментации используется искусственно созданный корпус текстов. Он создан на основе коллекции PostScience. В экспериментах используются искусственные документы для оценки, как это делалось во многих работах ранее [15; 16; 47]. В данном исследовании используются именно они, потому что сравнение с реальными текстами — сложная задача, так как не существует эталонных границ сегментов. Искусственные документы создаются методом объединения полнотекстовые документы из коллекции PostScience. Этот метод оказался более эффективным по сравнению с другими способами создания искусственных документов [15].

Данная глава организована следующим образом: сначала описываются принципы работы тематических моделей, методы интеграции тематической информации в процесс сегментации, а также экспериментальные результаты, подтверждающие эффективность предлагаемого решения по сравнению со случайным или структурным разбиением текста. Показывается, что этот подход позволяет повысить связность и точность перевода, минимизирует ошибки, связанные с неправильным определением контекста, и открывает новые возможности для дальнейшего развития методов машинного перевода длинных документов. Далее рассматривается применение полученного метода тематической сегментации к задаче информационного поиска. Результаты, изложенные в данной главе касательно тематической модели, использующей сегментную структуру документов, подробно изложены в работе соискателя [57]. Использование данного подхода в задаче информационного поиска — в работе [58], в которой соискатель отвечал за качество построения тематических сегментирующих моделей на основе предложенного в [15] способа.

В конце главы рассматривается подход к улучшению качества машинного перевода за счёт предварительной тематической сегментации текста.

5.1 Аддитивная регуляризация тематических моделей

Тематическая модель описывает коллекцию D с помощью конечного набора тем T. В ARTM [13] и в более базовой модели PLSA [48] распределение слов в документах моделируется как смесь тем:

$$p(w \mid d) = \sum_{t \in T} p(w \mid t) p(t \mid d), \quad d \in D,\ w \in W, \tag{5.1}$$

Модель параметризуется стохастическими матрицами Φ и Θ с элементами:

$$\varphi_{wt} = p(w \mid t), \quad \theta_{td} = p(t \mid d) \tag{5.1}$$

Тематическое моделирование можно также интерпретировать как задачу приближённой матричной факторизации $F \approx \Phi\Theta$. Решение задачи матричной факторизации не является единственным, поэтому, следуя подходу ARTM [13], следует учитывать дополнительные критерии для более точного определения матриц Φ и Θ. В частности, можно максимизировать взвешенную сумму логарифмической вероятности и некоторых аддитивных регуляризаторов R_i:

$$\sum_{d \in D} \sum_{w \in W} n_{dw} \ln \sum_{t \in T} \varphi_{wt} \theta_{td} + \sum \tau_i R_i(\Phi, \Theta) \to \max_{\Phi, \Theta}, \tag{5.2}$$

Регуляризаторы R_i накладывают дополнительные ограничения на параметры модели в зависимости от конкретной задачи. Коэффициенты регуляризации τ_i уравновешивают важность регуляризаторов и логарифмической правдоподобности. Если регуляризаторы не используются, описанная модель упрощается до PLSA [48]. Стационарная точка задачи 5.2 удовлетворяет системе уравнений, что позволяет использовать алгоритм максимизации ожидания в качестве метода итераций с фиксированной точкой. E-шаг этого алгоритма вычисляет вероятности отнесения слов к темам в контексте документа $p(t \mid d, w) \equiv p_{tdw}$. M-step использует эти вероятности для обновления матриц Φ и Θ.

5.2 Использование сегментной структуры документов для улучшения EM-алгоритма

Согласно 5.1, каждый документ представлен в виде «мешка слов». Аддитивные регуляризаторы обычно применяются на этапе M, и они также не могут делать никаких предположений о порядке слов. Однако на этапе E мы последовательно вычисляем вероятности p_{tdw} для каждой позиции в документе. Это означает, что можно делать дополнительные предположения о тематическом распределении слов, которые встречаются в одной и той же части текста. Согласно этим предположениям, значения p_{tdw} могут быть изменены, а затем использованы на этапе M алгоритма EM.

В реальных текстах авторы обычно излагают свои мысли последовательно. Поэтому можно ожидать, что в любом небольшом фрагменте текста будет всего несколько тем. Это можно сформулировать как предположение о разреженности тем в предложении $p(t|s)$, где мы определяем тему предложения как среднюю сумму распределений слов по темам:

$$p_{ts} \equiv p(t \mid s) = \frac{1}{n_s} \sum n_{sw} p_{tdw} \tag{5.2}$$

где n_s — длина предложения, а n_{sw} — количество вхождений слова в предложение.

Можно показать, что приведённое допущение о разреженности влияет на ptdw следующим образом:

$$\widetilde{p_{tdw}} = p_{tdw} \left(1 - \frac{\tau}{n_{dw}} \sum_{s \in S_d} \frac{n_{sw}}{n_s} \left(\frac{1}{p_{ts}} - \sum_{z \in T} \frac{p_{zdw}}{p_{zs}} \right) \right), \tag{5.3}$$

где S_d — множество всех предложений в документе.

Для данной формулы можно дать некоторую интерпретацию того, как она работает. Если p_{ts} для какого-то предложения близко к нулю, то $\frac{1}{p_{ts}}$ велико и результирующий знак для этого предложения будет отрицательным. Это означает, что вероятность соответствующей темы в слове уменьшится. Напротив, если p_{ts} близко к 1, этот член предложения может быть положительным, и вероятность соответствующей темы увеличится. Короче говоря, каждый член суммы в этой формуле приближает распределение p_{tdw} к основным темам предложений, в которых встречается слово.

Стоит отметить, что $\widetilde{p_{tdw}}$ может не определять распределение по темам, поскольку значения могут быть отрицательными, но это не нарушает работу алгоритма EM. Параметр τ в формуле определяет силу влияния субъектов предложения на распределение слов по темам. Рассмотрим идею разреженности субъектов текстовых фрагментов и применим ее для задачи тематической сегментации. Предполагается, что любой текст состоит из сегментов, которые можно представить в виде небольшого количества тем. Предполагается, что темы остаются неизменными в пределах каждого сегмента. Предполагается, что два последовательных сегмента имеют мало пересекающихся тем. Теперь постепенно будем определять границы сегментов в процессе обучения модели. На первой итерации EM-алгоритма будем брать предложения в качестве начального приближения для сегментов. Затем на каждом этапе E будем находить темы сегментов и объединять последовательные сегменты, если у них одна и та же тема. Уравнение 5.3 в этом методе применяется к сегментам, которые были созданы на текущей итерации.

5.3 Оценка качества тематической модели

Для оценки качества тематической модели будем учитывать два фактора: качество сегментации тематической модели и разреженность внутри документа. Качество сегментации показывает способность модели восстанавливать границы тем и семантические изменения в тексте. Для оценки будем использовать искусственно созданный корпус текстов. Он предоставит нам золотой стандарт для определения границ сегментов в документах. Чтобы проверить, насколько границы золотого стандарта совпадают с предполагаемыми границами, будем использовать показатели Pk и WindowDiff, как это было сделано в предыдущей работе [15]. Разреженность внутри документа показывает способность модели описывать семантические сегменты с помощью минимально возможного количества тем. Разреженность тем в сегментах подразумевает разреженность всего документа, поэтому для оценки мы будем использовать среднюю разреженность Θ-матрицы.

Определение границ сегментов с помощью тематических моделей. Для всех тематических моделей мы будем использовать специальный алгоритм сег-

ментации тем для определения границ сегментов. Этот метод был применён в алгоритме TopicTiling [15] и показал хорошие результаты по сравнению с другими алгоритмами сегментации. Идея этого метода заключается в вычислении сходства между левым и правым окнами для каждого окончания предложения. Окончания предложений с наименьшими значениями этого сходства считаются кандидатами на роль границ сегментов. Затем к функции сходства применяется некоторое сглаживающее преобразование для получения так называемого показателя глубины. Кандидаты, у которых показатель глубины превышает определенный порог, выбираются в качестве конечных границ сегмента. Показатель глубины также может быть интерпретирован как вероятность того, что граница в соответствующем предложении заканчивается. Наша версия алгоритма сегментирования отличается от оригинальной, поскольку мы используем темы предложений для вычисления сходства между окнами предложений. В оригинальной версии авторы использовали идентификаторы тем, присвоенные словам во время вывода.

Показатели качества сегментации Pk и WindowDiff. Pk-мера использует скользящее окно длиной в k токенов, которое перемещается по тексту для расчёта штрафов за сегментацию. Для каждой пары слов, находящихся на расстоянии друг от друга, проверяется, относятся ли они к одному и тому же сегменту или к разным сегментам. Это делается отдельно для границ золотого стандарта и границ предполагаемых сегментов. Если золотой стандарт и предполагаемые сегменты не совпадают, добавляется штраф в размере 1. Наконец, частота ошибок вычисляется путём нормализации штрафа по количеству пар. Значение, близкое к 0, указывает на идеальное качество сегментации при оценке. Значение параметра k равно половине количества токенов в документе, делённой на количество сегментов, указанное в золотом стандарте. Недостатком показателя Pk является то, что он не учитывает количество сегментов между парой слов. WindowDiff — это усовершенствованная версия Pk: в ней учитывается количество сегментов между парой слов. Затем сравнивается количество сегментов в золотом стандарте и в оценённых сегментах. Если количество сегментов не совпадает, к штрафу добавляется 1, и результат снова делится на количество пар, чтобы получить коэффициент ошибок в диапазоне от 0 до 1 [15].

Описание искусственно созданного корпуса текстов. В качестве основы для созданной коллекции мы используем корпус PostScience. Мы применяем лемматизацию, удаляем стоп-слова и все документы, содержащие более 200 или менее 10 предложений. Затем мы составляем искусственные документы, объеди-

няя полные исходные документы. Как упоминалось в [15], использование полных документов делает корпус более реалистичным по сравнению со случаями, когда объединяются только фрагменты документов. Чтобы избежать повторения тем в последовательных сегментах, мы создаем простую тематическую модель на основе набора данных PostScience и используем только документы с разными темами для последовательных сегментов. Кроме того, мы используем только те документы, вероятность того, что одна тема превысит порог в 0,8. Все это позволяет нам предположить, что границы сегментации по золотому стандарту также являются тематическими границами. Количество сегментов в документе варьируется от 2 до 4. Итоговое количество документов в сгенерированном корпусе равно 700.

Настройка экспериментов. Во всех представленных экспериментах для построения тематической модели использовалась библиотека с открытым исходным кодом BigARTM [21]. Для поиска оптимальных значений параметров мы используем 5-кратную перекрестную проверку на обучающем подмножестве. Оно включает 500 искусственных документов. Для оценки используется метрика WindowDiff. Опишем все параметры, которые мы изучаем:

- I - количество итераций в EM-алгоритме. Этот параметр тесно связан с переобучением и сходимостью модели.
- α — сила регуляризатора разреженности Theta. Изменение этого параметра исследуется, чтобы выявить влияние обычной разреженности на качество сегментации в тематических моделях.
- τ_1 — параметр τ в уравнении 5.3, который используется, когда границы сегментов совпадают с границами предложений.
- τ_2 — параметр τ уравнения 5.3, который используется при итеративном вычислении границ сегментов путем объединения последовательных сегментов.
- w — размер окна, который используется в алгоритме сегментации для вычисления окончательных границ.

Результаты настройки параметров приведены в таблице 16. После определения оптимального количества итераций мы определяем структуру процесса обучения. В течение первых 5 итераций тематическая модель работает без каких-либо регуляризаторов. На 5-й итерации начинает работать регуляризатор разреженности Θ. Для последних 25 итераций мы применяем уравнение 5.3 на каждом E-шаге алгоритма. Такая структура обучения необходима, поскольку для использования сегментированных объектов требуется сходимость тематической

Таблица 16 — Результаты экспериментов по оценке оптимальных гиперпараметров тематической модели.

Parameter	**Optimal value**	**WindowDiff**
I	40	0.253
α	0.2	0.248
τ_1	0.1	0.242
τ_2	11	0.232

модели. Как видно из таблицы 16, оптимальное значение параметра α очень мало. Это означает, что сильное влияние регуляризатора разреженности Θ снижает способность тематической модели восстанавливать границы сегментов. Можно также отметить, что итеративное объединение сегментов дает высокий коэффициент τ_2 в уравнении 5.3, в то время как стратегия с фиксированными границами предложений сохраняет этот коэффициент низким. Это означает, что создание разреженных небольших сегментов, таких как предложения, нежелательно. Это можно объяснить тем, что небольшие предложения в большей степени зависят от тематики отдельных слов и их тематика более нестабильна. Оптимальным значением параметра w было 11. Это может быть объяснено тем, что самые короткие сегменты золотого стандарта в нашей коллекции имеют размер 10 предложений. Таким образом, при реальной выборке мы рекомендуем установить этот параметр равным длине самого короткого сегмента.

5.4 Эксперименты по улучшению качества сегментации

Все модели с оптимальными значениями параметров были протестированы на тестовых документах из нашего искусственного набора данных. Количество тестовых документов — 200. Обучающие документы использовались только для построения тематической модели. Окончательные результаты сегментации представлены в таблице 17 и сравниваются с моделью TopicTiling из [15]. Для базовой модели TopicTiling воспроизводится исходная оценка сходства между окнами на основе распределения тем LDA. В наших моделях используются темы предложений для вычисления сходства. PLSA + разреженность Θ – это модель, которая использует регуляризатор Θ разреженности. PLSA + SentenceSparse – это модель,

которая применяет уравнение 5.3, на уровне предложений. PLSA + SegmentSparse – это модель с итеративным объединением сегментов. Во всех использованных моделях используются подобранные гиперпараметры.

Таблица 17 — Сравнение качества сегментации различных подходов.

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

Как видно, предложенный метод с регуляризацией 5.3 и итеративным объединением сегментов обеспечивает наилучшее качество сегментации. Сравнивая SegmentSparse и SentenceSparse улучшение составляет 0.4% по метрике P_k, но в случае задачи перевода — даже разовая неверная расстановка границ сегмента может привести к разрыву контекста и последующим ошибкам перевода. Кроме того, модели TopicTiling и PLSA + разреженность Θ, построенные на основе модели «мешок слов», показывают наихудшие результаты в сегментации. Чтобы лучше изучить эффективное количество тем в моделях, приведены сравнения уровня разреженности матрицы Θ в таблице 18. Модель SegmentSparse сокращает среднее количество тем почти в 3 раза. Без использования уравнения 5.3 такой результат в плане разреженности мог быть достигнут только за счёт снижения качества сегментации.

Таблица 18 — Результаты расчёта разреженности для различных моделей

Название модели	**Соотношение ненулевых значений в Θ**
PLSA + разреженность Θ	4%
PLSA + SentenceSparse	1.8%
PLSA + SegmentSparse	1.5%

Теперь рассмотрим, как темы распределяются по документу. На рис. 5.1 мы обозначили доминирующие темы для каждого слова в последовательном тексте цветом. Модель PLSA + разреженность Θ выявляет в тексте только два семантических сегмента, а во втором сегменте все темы перемешаны. В то же время модель SegmentSparse выделяет все три сегмента и делает их тематически разными. Желтая тема присутствует как во втором, так и в третьем сегменте, поэтому

не будем выделять ее слова в данных фрагментах текста. Также обратим внимание, что вторая модель использовала 32 темы, в то время как для первой модели было достаточно 9 тем. Таким образом, в этом документе модель SegmentSparse превзошла PLSA + разреженность Θ как по разреженности, так и по оценке границ сегментов.

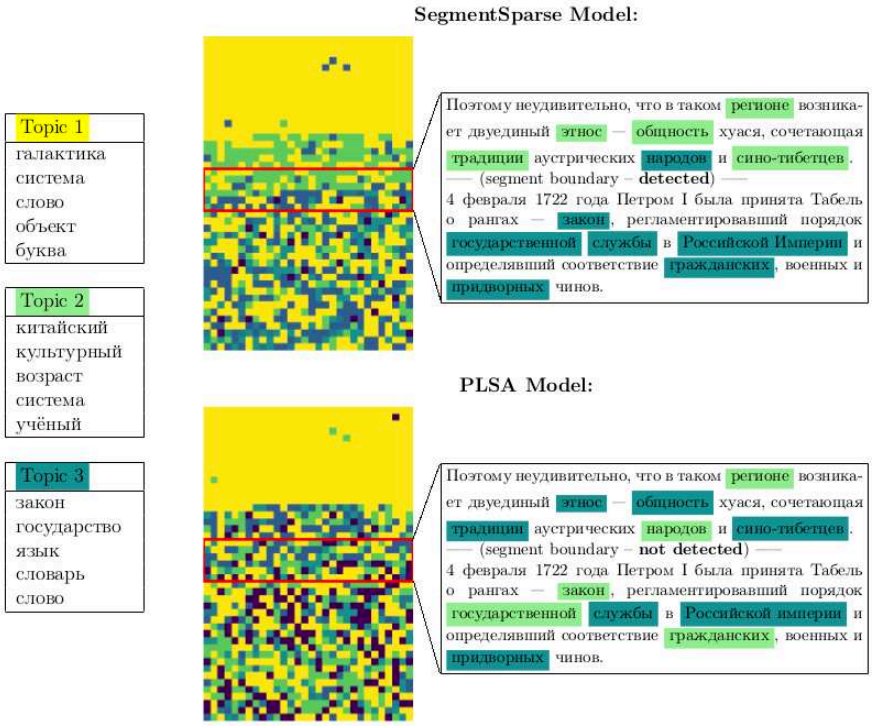


Рисунок 5.1 — Визуализация моделей PLSA + Θ sparsity reg. (внизу) и SegmentSparse (вверху), применённых к одному и тому же тестовому документу. Слова на рисунке представлены пикселями, которые следуют друг за другом слева направо и сверху вниз. Фрагмент текста, в котором модель PLSA дала сбой, отмечен красным.

Теперь представим результаты применения модели SegmentSparse к документам из исходной коллекции PostScience. По-прежнему будет использоваться модель, обученная на искусственных документах. На рис. 5.2 представлен исходный документ, в котором модель SegmentSparse обнаружила два разных сегмента. Первый тематический раздел содержит историческую информацию, а второй — сведения о природных особенностях региона. Таким образом, такая сегментация документа оправдана.

... Казанская губерния, наоборот, вошла по просьбе Коржинского. Это интересная историческая ситуация, но постепенно, уже к 20-м годам XX века средняя Россия охватывает территорию от Ярославской и Костромской губерний на севере до Воронежской и Саратовской на юге. Вот эта вся территория находится в пределах европейской части России на левобережье Волги. В среднем, по некоторым подсчётам, природная флора этого региона насчитывает примерно 4,5 тысячи видов, очень немного. Для сравнения флора Турции, которая меньше по площади, включает больше 15 тысяч видов. Связана бедность флоры, с одной стороны, с тем, что это равнинная территория...

Рисунок 5.2 — Тематическое представление исходного документа из PostScience, в котором модель SegmentSparse обнаружила 2 сегмента.

5.5 Использование тематической сегментации для улучшения качества информационного поиска

В предыдущем разделе продемонстрировано качество предложенного соискателем метода на задаче сегментации. Прикладные задачи в области обработки и понимания текстов естественного языка часто используют сегментацию как часть более сложной системы. Цель данного исследования — показать пользу тематической сегментации в задаче перевода, но в качестве дополнительной верификации полезности предложенного подхода рассмотрим его применение в контексте информационного поиска.

Цель информационного поиска (ИП) — сопоставить запрос с релевантными документами. Для этого можно использовать различные методы. Несмотря на успех глубоких нейронных сетей в решении множества задач обработки текстов естественного языка, в области ИП традиционный подход, основанный на терминах, по-прежнему остаётся популярным: ранжирование BM-25 [49] в сочетании с расширением запроса и, возможно, некоторыми дополнительными эвристическими методами по сей день является стандартной практикой в ИП.

Некоторые исследователи [50] обсуждали применение сегментации текста в задачах информационного поиска, но результаты оказались довольно спорными, и текущее состояние этой области остаётся неясным. В области сегментации текста существуют различные подходы к разделению текста на семантически однородные блоки. Среди них преобладают методы сегментации с использованием лексических цепочек и тематического моделирования.

Интуитивное понимание того, что сегментация текста улучшит качество поиска, можно выразить абстрактно: чем лучше моделируется язык и его особен-

ности, тем лучше должна работать любая модель или алгоритм, применяемые к языковым данным. Предполагается, что текст на естественном языке — это не хаотичный набор тем, а тематически структурированная единица: в реальной речи сначала раскрывается одна тема, затем происходит переход к другой, и таким образом темы следуют одна за другой. В вычислительных задачах может быть полезно смоделировать эту структуру сегментов. Сегментация по темам позволила бы разделить семантически отличимые части текстов и выявить более связные и однородные сегменты, с которыми легче работать.

Следующий аргумент основан на инженерных соображениях. В стандартной задаче поиска по запросу пользователя обычно интересуют не целые документы, выдаваемые информационно-поисковой системой, а только короткие абзацы, наиболее соответствующие заданному запросу. Таким образом, цель поиска по запросу состоит не только в том, чтобы найти релевантные документы, но и в том, чтобы выбрать наиболее релевантные части этих документов [50].

В данном разделе будет представлено описание информационно-поисковой системы, когда пользователи могут запрашивать не только по коротким ключевым фразам, но и по примерам документов. В качестве метода сегментации документов для поиска используется подход, основанный на тематическом моделировании и описанный в разделе 5.2.

Подробные результаты исследования можно увидеть в работе [58]. Соискатель разработал в рамках этой работы процесс построения тематической модели сегментации.

Как и было описано ранее, тематическая сегментация выполняется в два этапа:

1. Сначала строится тематическая модель на основе так называемого допущения о разреженности, согласно которому слова в сегменте должны относиться к одному и тому же небольшому набору тем. Это допущение позволяет ввести дополнительный регуляризатор на этапе E-шага, который используется для построения модели. Границы сегментов затем оцениваются постепенно.
2. Во-вторых, после построения тем по документу, созданным на предыдущем шаге, применяется алгоритм TopicTiling. Для каждой границы предложения алгоритм рассматривает левое и правое окна определённой длины n и вычисляет меру сходства между ними. Затем вычисленные значения сходства сглаживаются, и в качестве границ сегментов пред-

лагаются кандидаты с итоговым показателем сходства, превышающим выбранное пороговое значение.

Для построения границ сегментов пороговое значение вычисляется по следующей формуле:

$$\text{порог} = E[ds] - \alpha\sqrt{E[ds^2] - (E[ds])^2},$$

где $E[ds]$ — среднее значение показателей глубины, $\sqrt{E[ds^2] - (E[ds])^2}$ — стандартное отклонение показателей глубины, а α — коэффициент, который можно изменять для настройки детализации сегментации. Значение α по умолчанию равно 0.5.

Для оценки полезности сегментации приведем краткое описание схемы поиска, которая была разработана не в рамках темы данного диссертационного исследования. Предлагается использовать следующий алгоритм поиска. После предварительной обработки текста целые документы делятся на квазиабзацы — относительно небольшие семантически связанные части. На этом этапе используется метод тематической сегментации. Затем создаётся инвертированный индекс от абзацев к документам, чтобы обеспечить дальнейшее сопоставление абзацев с документами. Полученные абзацы векторизуются. На рисунке 5.3 показан описанный выше процесс.

Данные. В качестве обучающих данных используется общедоступная коллекция статей из репозитория arXiv. Этот набор данных используется для обучения модели сегментации. Для проверки качества поиска используется автоматически сгенерированные наборы данных в виде триплетов по методологии, описанной в [51]. Триплеты содержат в себе статью□запрос, релевантную запросу статью и нерелевантную запросу статью. Релевантная статья имеет ту же тематику, что и статья-запрос. Нерелевантная статья не имеет пересечений по темам со статьей из запроса.

Всего было собрано около 140 000 статей. Обучающий набор данных включает 95 000 статей, а тестовый набор — 45 000 уникальных статей, которые формируют 15 715 триплетов.

Также тексты проходили несколько этапов предобработки, включая токенизацию и лемматизацию. Кроме того удалялись короткие строки, а также математические символы и формулы.

В работе [58] рассматривается несколько моделей сегментации, однако в данном исследовании сосредоточимся на оценке влияния тематической сег-

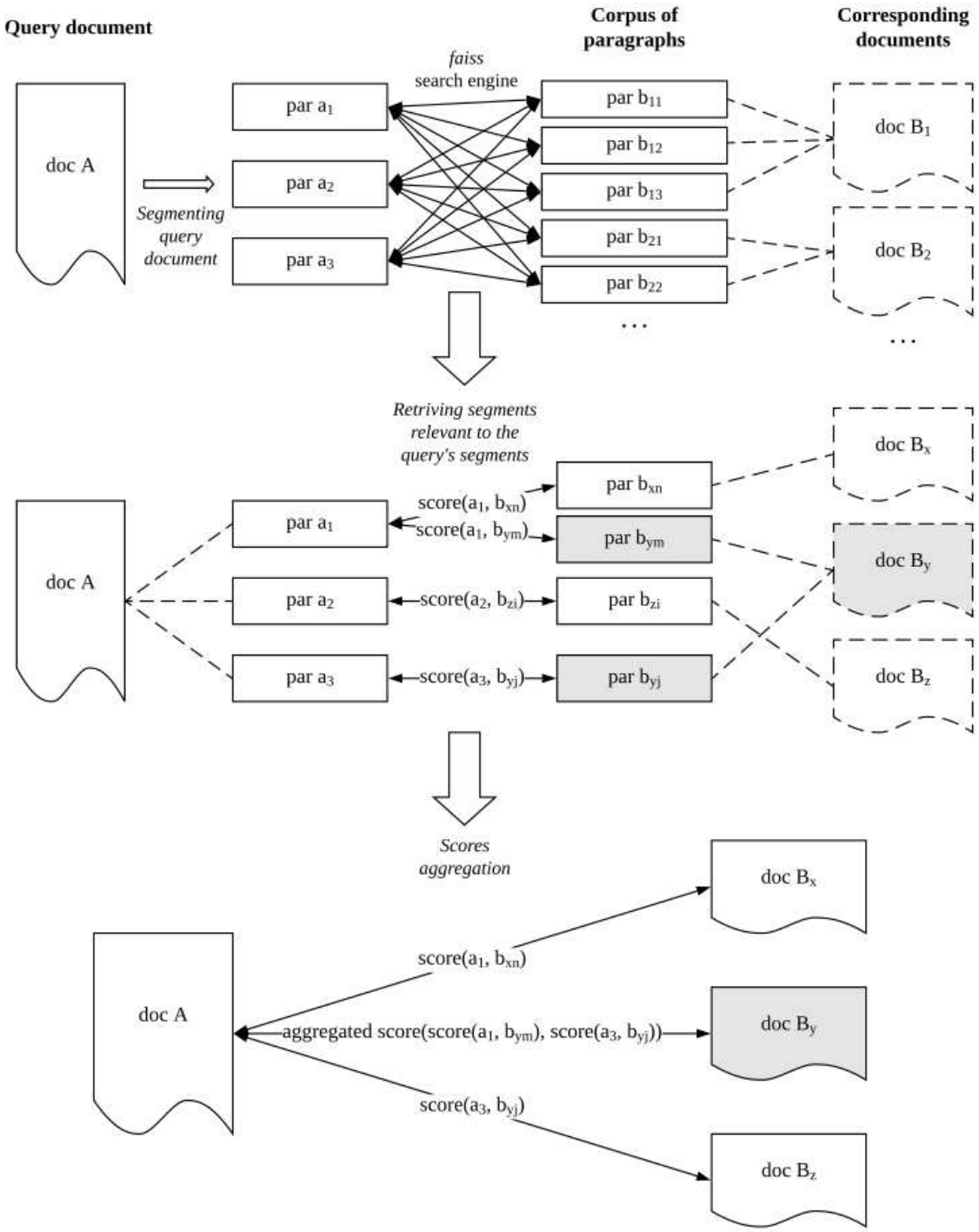


Рисунок 5.3 — Визуализация схемы поиска, основанной на сегментации документов

ментации на результаты поиска. Для этого необходимо описать, каким образом оценивается качество поиска.

Оценка качества информационного поиска. В данной задаче оценка качества поиска и сравнение базовых моделей будет проводиться с помощью точности. Данная метрика выбрана из-за синтетической природы данных, используемых для оценки качества. Точность здесь определяется как $1 - \frac{\text{Количество инверсий}}{\text{Общее количество троек}}$,

где количество инверсий — это абсолютное количество случаев, когда рейтинг неправильного документа выше, чем рейтинг нужного документа.

В сравнении также приведено сравнение двух разных степеней детализации сегментаций: более детализированная сегментация с порогом $\alpha = 0.5$ и менее детализированная с порогом $\alpha = 0.3$.

Результаты экспериментов с сегментацией в задаче информационного поиска. Результаты эксперимента описаны в таблице 19. Жирным шрифтом выделены наилучшие конфигурации для каждой строки. Как видно, в большинстве случаев тематическая сегментация улучшает качество поиска. Различные подходы к уровню детализации сегментаций дают немного отличающиеся результаты. Более того, похоже, что различные стратегии агрегирования существенно влияют на показатели точности. Предварительное наблюдение: агрегирование оценок трёх наиболее подходящих сегментов даёт наилучший результат, как и в случае с большинством базовых моделей. Также выяснилось, что более тонкая сегментация немного лучше подходит для конкретной задачи. Поэтому было бы разумно изучить методы создания более эффективных сегментов для достижения лучших результатов в последующих задачах информационного поиска и протестировать более сложные стратегии агрегирования, включающие, например, взвешивание семантической полезности абзацев. Примечательно, что модель на основе ARTM лучше работает с целыми текстами. Этот результат не удивителен, поскольку тематическое моделирование по своей природе создаёт агрегированное тематическое представление документа, которое лучше, чем любое простое усреднение его сегментов. При этом видно, что тематическая модель, построенная с учетом сегментной структуры документов, даёт хорошее качество сегментации.

5.6 Улучшение качества перевода

Перевод длинных документов — трудоёмкая и сложная задача, которая требует от автоматических моделей перевода не только знания языков, но и умения сохранять контекстную связность текста. В реальном мире, задача контекстного перевода текста произвольной длины решается значительно хуже чем задача перевода на уровне предложений [17]. Во многом по этой причине в задаче машинного перевода преуспевают модели перевода основанные на больших язы-

Таблица 19 — Результаты экспериментов с сегментацией текстов в задаче информационного поиска

Model	**Точность без сегментации**	**Точность с сегментацией**
ARTM	**0.817**	0.783
sent2vec	0.770	**0.809**
fastText	0.751	**0.785**
doc2vec	**0.814**	0.785
avW2V	0.817	**0.824**
avnormW2V	0.580	**0.620**
avGloVe	0.779	0.779
avnormGloVe	0.573	**0.609**
avfastText	0.662	**0.746**

ковых моделях [59], ведь последние на этапе предобучения учатся на текстах значительной длины и лучше используют контекстные связи на этапе генерации.

Сложность задачи контекстного перевода текстов произвольной длины усугубляется необходимостью дробить текст на смысловые блоки. Случайное разбиение длинного документа на сегменты для перевода может привести к потере смысловых связей и ухудшению качества конечного перевода.

В данной главе рассматривается метод текстовой сегментации на основе тематических моделей ARTM [13], который позволяет разбивать длинные документы на однородные по тематике сегменты. Необходимыми свойствами обладают рассматриваемые в данной главе тематические модели, которые используют сегментную структуру документов при обучении [**segment-tm**]. Подробное описание таких тематических моделей можно увидеть в разделе 5.2.

Использование такого подхода в задаче перевода обладает некоторыми преимуществами. Во-первых, тематические модели достаточно легковесны и могут быть предобучены на общем домене. Соответственно, использование сегментации текстов с помощью тематических моделей может быть предпочтительнее вычислительно затратных моделей по типу BERT [9]. Во-вторых, предложенный соискателем метод гранулярности и интерпретируемости тематических моделей открывает возможности для более контролируемой настройки сегментации, часто необходимой профессиональным переводчикам при работе с CAT-инструментами.

В данном разделе будет продемонстрирована эффективность использования текстовой сегментации на основе ARTM в задаче перевода и показано преимущество данного метода перед случайным разбиением текста на сегменты. В главе будут описаны этапы подготовки данных и применения сегментации и перевода к документам из WMT22 [52], а также представлены результаты экспериментального исследования, подтверждающего эффективность предложенного подхода.

5.6.1 Условия экспериментов

Для построения тематической модели использовалась библиотека BigARTM [21] с использованием процедуры постобработки E-шага, описанной в разделе 5.2.

Обучение тематической модели проводилось на коллекции из 5000 документов англоязычной википедии SimpleWiki. При построении модели документы коллекции были лемматизированы, а также очищены от пунктуации и стоп-слов.

Обучение тематической модели проводилось с использованием регуляризаторов разреживания матрицы Θ с коэффициентом -0.1 на протяжении 20 эпох. До этого модель училась без использования регуляризаторов на протяжении 5 эпох. Далее, заключительные 5 эпох обучения проводились с использованием сегментирующего регуляризатора с весом 15 и с порогом объединения сегментов, равным 0.5. При использовании тематической модели для сегментации все настройки были сохранены. Количество итераций на документе при применении модели равнялось пяти.

Оценка качества перевода проводилась с помощью метрики BLEU [18] на наборе WMT22 [52]. Тестовый набор этого года представлял из себя документы, а не предложения. Для их перевода важен учет контекстных связей. Документы WMT22 были объединены в один документ, что моделирует ситуацию перевода потока, когда границы документов при переводе неизвестны.

Для перевода сегментов использовалась модель Qwen3-14B через облако [53]. Она обладает приемлемым качеством перевода с английского на русский и, как и многие языковые модели, способна переводить тексты из многих предложений с учетом контекстных связей.

5.6.2 Результаты экспериментов с контекстным переводом

В экспериментах с сегментацией сравнивается качество перевода при разбиении текстов WMT22 на сегменты случайным образом и с использованием тематической модели. Количество сегментов при сегментации обоими способами было выбрано одинаковым, исходя из того, сколько реально документов использовалось в WMT22. Качество перевода для разных способов сегментации текста можно увидеть в таблице 20.

Таблица 20 — Качество перевода при использовании различных алгоритмов сегментации

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

Как показали результаты экспериментов, использование тематической сегментации текста перед автоматическим переводом даже при неизменном количестве сегментов позволяет добиться улучшения метрики качества перевода BLEU по сравнению со случайным разбиением текста на части. Несмотря на то, что абсолютный прирост по метрике BLEU оказался относительно небольшим [34], на практике семантически мотивированные (тематически однородные) сегменты обеспечивают более точное сохранение значимых смысловых связей при переводе длинных документов, способствуют уменьшению локальных смысловых и грамматических ошибок, а также делают перевод более связным с точки зрения восприятия целостного текста на целевом языке. Для демонстрации этого эффекта в таблице 21 можно увидеть примеры переводов, в которых более семантически точное деление текстов на сегменты позволило лучше уловить смысл текстов при переводе по сравнению со случайной сегментацией.

В первом из приведенных примеров можно увидеть, что модель перевода с семантической сегментацией благодаря увиденному контексту смогла точнее передать смысл исходного текста с сохранением гладкости перевода, тогда как при случайной сегментации в результате более дословного перевода получилась смысловая ошибка. Во втором примере видно, что модель перевода с семантической сегментацией смогла верно перевести анафору благодаря более качественной расстановке границ сегментов.

Таблица 21 — Примеры переводов документов WMT22 с использованием случайной и семантической сегментаций. Символ | — означает границу сегмента.

Входной текст	Перевод модели со случайной сегментацией	Перевод модели после сегментации PLSA+SegmentSparse
«We’re going to get through this together and the federal government is not going to walk away», he said. «This is one of those times when we aren’t Democrats or Republicans.»	«Мы преодолеем это вместе, и федеральное правительство не уйдёт в отставку», — сказал он. | «Это один из тех многих моментов, когда мы не демократы и не республиканцы»	«Мы преодолеем это вместе, и федеральное правительство не уйдёт в отставку», — сказал он. «Сейчас мы не демократы и не республиканцы»
This technology is currently before commercialization. However, it can be used as an auxiliary in current situations where the PCR test for Omicron has not been developed.	В настоящее время эта технология ещё не коммерциализирована. | Однако его можно использовать в качестве вспомогательного средства сейчас, когда тест ПЦР на Омикрон ещё не разработан	В настоящее время эта технология ещё не коммерциализирована. Однако её можно использовать в качестве вспомогательной сейчас, когда ПЦР-тест на Омикрон ещё не разработан

5.7 Выводы

Данные экспериментов подтверждают, что интеграция тематического моделирования на этапе предварительной обработки текстов оказывает положительное влияние на итоговое качество контекстного машинного перевода с точки зрения метрики BLEU [18]. Такой подход может быть особенно полезен для сложных по структуре и содержанию документов, где границы между тематическими фрагментами не совпадают с формальными абзацами или предложениями. Кроме того предложенная соискателем методика может быть эффективно использована как часть комплексных пайплайнов перевода для профессиональных систем локализации, корпоративного документооборота и автоматизации работы с большими

текстовыми массивами, так как задача перевода длинных текстов разбивается на поддающиеся человеческой интерпретации этапы сегментации и перевода.

В основе предложенного соискателем алгоритма сегментации лежит алгоритм постобработки E-шага, реализованный в библиотеке [21]. Предложенный подход растит качество сегментации на корпусе PostScience относительно сегментаций, построенных на базовых тематических моделях. Эффективность предложенного метода сегментации также проверена в задаче информационного поиска.

Таким образом, тематическая сегментация служит возможным инструментом улучшения качества итогового перевода. Правильно построенные сегменты позволяют лучше учитывать контекстные связи и обеспечивать сохранение целостности и связности смысла при переводе длинных и разнородных по структуре документов.

Заключение

Основные положения, выносимые на защиту:

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 [18] 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 [18], 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 [13]. 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 [18].

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.

Список литературы

1. *Bahdanau, D.* Neural Machine Translation by Jointly Learning to Align and Translate [Текст] / D. Bahdanau, K. Cho, Y. Bengio // CoRR. — 2015. — abs/1409.0473.
2. Attention is All You Need [Text] / A. Vaswani, N. Shazeer, N. Parmar, [et al.] // Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17). — Red Hook, N.Y, 2017. — P. 6000—6010.
3. *Provilkov, I.* BPE-Dropout: Simple and Effective Subword Regularization [Текст] / I. Provilkov, D. Emelianenko, E. Voita // Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics / под ред. D. Jurafsky [и др.]. — Online : Association for Computational Linguistics, 07.2020. — С. 1882—1892.
4. FlowSeq: Non-Autoregressive Conditional Sequence Generation with Generative Flow [Текст] / X. Ma [и др.] // Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP) / под ред. K. Inui [и др.]. — Hong Kong, China : Association for Computational Linguistics, 11.2019. — С. 4282—4292.
5. Latent-Variable Non-Autoregressive Neural Machine Translation with Deterministic Inference Using a Delta Posterior [Текст] / R. Shu [и др.] // Proceedings of the AAAI Conference on Artificial Intelligence. — 2020. — Апр. — Т. 34. — С. 8846—8853.
6. Investigating Backtranslation in Neural Machine Translation [Text] / A. Poncelas [et al.] // Proceedings of the 21st Annual Conference of the European Association for Machine Translation / ed. by J. A. Pérez-Ortiz [et al.]. — Alicante, Spain, 2018. — May. — P. 269—278.
7. Dual Learning for Machine Translation [Текст] / Y. Xia, D. He, T. Qin [и др.] // Proceedings of the 30th International Conference on Neural Information Processing Systems (NIPS'16). — Red Hook, N.Y., 2016. — С. 820—828.

8. Reducing Word Omission Errors in Neural Machine Translation: A Contrastive Learning Approach [Text] / Z. Yang, Y. Cheng, Y. Liu, [et al.] // Proc. 57th Annual Meeting of the Association for Computational Linguistics. — Florence, 2019. — P. 6191—6196.

9. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding [Text] / J. Devlin, M. Chang, [et al.] // Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. — Minneapolis, Minnesota, 2019. — P. 4171—4186.

10. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension [Text] / M. Lewis, Y. Liu, [et al.] // Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. — Seattle, 2020. — P. 7871—7880.

11. mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer [Text] / L. Xue [et al.] // Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies / ed. by K. Toutanova [et al.]. — Online, 2021. — June. — P. 483—498.

12. Context-Aware Neural Machine Translation Learns Anaphora Resolution [Текст] / E. Voita [и др.] // Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) / под ред. I. Gurevych, Y. Miyao. — Melbourne, Australia : Association for Computational Linguistics, 07.2018. — С. 1264—1274.

13. *Vorontsov*, *K.* Additive Regularization of Topic Models [Text] / K. Vorontsov, A. Potapenko // Machine Learning. — 2014. — Vol. 101. — P. 1—21.

14. *Blei*, *D. M.* Latent Dirichlet Allocation [Text] / D. M. Blei, A. Y. Ng, M. I. Jordan // Journal of Machine Learning Research. — 2003. — Vol. 3. — P. 993—1022.

15. *Riedl*, *M.* Text Segmentation with Topic Models [Text] / M. Riedl, C. Biemann // Journal for Language Technology and Computational Linguistics (JLCL). — 2012. — Vol. 27. — P. 47—69.

16. *Du*, *L.* Topic Segmentation with a Structured Topic Model [Text] / L. Du, W. Buntine, M. Johnson // Proceedings of NAACL-HLT 2013. — 2013. — P. 190—200.

17. Challenges in Context-Aware Neural Machine Translation [Text] / L. Jin [et al.] // Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing / ed. by H. Bouamor, J. Pino, K. Bali. — Singapore, 2023. — Dec. — P. 15246—15263.

18. Bleu: A Method for Automatic Evaluation of Machine Translation [Text] / K. Papineni, S. Roukos, T. Ward, [et al.] // Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. — 2002. — P. 311—318.

19. *Sellam*, *T.* BLEURT: Learning Robust Metrics for Text Generation [Text] / T. Sellam, D. Das, A. Parikh // Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics / ed. by D. Jurafsky [et al.]. — Online, 2020. — July. — P. 7881—7892.

20. COMET: A Neural Framework for MT Evaluation [Text] / R. Rei [et al.] // Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP) / ed. by B. Webber [et al.]. — Online, 2020. — Nov. — P. 2685—2702.

21. Bigartm: Open-Source Library for Topic Modeling of Big Text Collections [Text] / K. Vorontsov [et al.] // Analytics and Data Management in Areas with Intensive Use of Data. DAMDID/RCDL'2015. — Obninsk, 2015. — P. 28—36.

22. *Stahlberg*, *F.* Neural Machine Translation: A Review [Текст] / F. Stahlberg // J. Artific. Intelligence Res. — 2020. — № 69. — С. 343—418.

23. *Hochreiter*, *S.* Long Short-Term Memory [Текст] / S. Hochreiter, J. Schmidhuber // Neural Computation. — 1997. — Т. 9, № 8. — С. 1735—1780.

24. *Kingma*, *D.* Adam: A Method for Stochastic Optimization [Text] / D. Kingma, J. Ba // 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, May 7-9, 2015. Conference Track Proceedings. — 2015.

25. Findings of the 2017 Conference on Machine Translation (WMT17) [Text] / O. Bojar, R. Chatterjee, C. Federmann, [et al.] // Proceedings of the Second Conference on Machine Translation. Volume 2: Shared Task Papers. — 2017.

26. Iterative Back-Translation for Neural Machine Translation [Text] / V. C. D. Hoang [et al.] // Proceedings of the 2nd Workshop on Neural Machine Translation and Generation / ed. by A. Birch [et al.]. — Melbourne, Australia, 2018. — July. — P. 18—24.

27. Scaling Laws for Neural Language Models [Текст] / J. Kaplan, S. McCandlish, T. Henighan [и др.]. — 2020. — arXiv: 2001.08361 [cs.CL].

28. *Shaw*, *P.* Self-Attention with Relative Position Representations [Text] / P. Shaw, J. Uszkoreit, A. Vaswani // Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers). — 2018. — P. 464—468.

29. *Barbaresi*, *A.* Trafilatura: A Web Scraping Library and Command-Line Tool for Text Discovery and Extraction [Текст] / A. Barbaresi // Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations / под ред. H. Ji, J. C. Park, R. Xia. — Online : Association for Computational Linguistics, 08.2021. — С. 122—131.

30. *Esplà-Gomis*, *M.* Bitextor: a Free/Open-source Software to Harvest Translation Memories from Multilingual Websites [Текст] / M. Esplà-Gomis // Beyond Translation Memories: New Tools for Translators Workshop. — Ottawa, Canada, 8 26-30.2009.

31. Bifixer and Bicleaner: two open-source tools to clean your parallel data [Text] / G. Ramírez-Sánchez [et al.] // Proceedings of the 22nd Annual Conference of the European Association for Machine Translation / ed. by A. Martins [et al.]. — Lisboa, Portugal, 2020. — Nov. — P. 291—298.

32. ParaCrawl: Web-Scale Acquisition of Parallel Corpora [Text] / M. Bañón, P. Chen, B. Haddow, [et al.] // Proc. 58th Annual Meeting of the Association for Computational Linguistics. — Seattle, 2020. — P. 4555—4567.

33. *Koehn*, *P.* Statistical Significance Tests for Machine Translation Evaluation [Текст] / P. Koehn // Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing / под ред. D. Lin, D. Wu. — Barcelona, Spain : Association for Computational Linguistics, 07.2004. — С. 388—395.

34. Navigating the Metrics Maze: Reconciling Score Magnitudes and Accuracies [Текст] / T. Kocmi [и др.] // Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) / под ред. L.-W. Ku, A. Martins, V. Srikumar. — Bangkok, Thailand : Association for Computational Linguistics, 08.2024. — С. 1999—2014.

35. *Vanmassenhove, E.* Getting Gender Right in Neural Machine Translation [Текст] / E. Vanmassenhove, C. Hardmeier, A. Way // Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing / под ред. E. Riloff [и др.]. — Brussels, Belgium : Association for Computational Linguistics, 10-11.2018. — С. 3003—3008.

36. Lost in Literalism: How Supervised Training Shapes Translationese in LLMs [Текст] / Y. Li [и др.] // Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) / под ред. W. Che [и др.]. — Vienna, Austria : Association for Computational Linguistics, 07.2025. — С. 12875—12894.

37. Diverse Beam Search: Decoding Diverse Solutions from Neural Sequence Models [Text] / A. K. Vijayakumar, M. Cogswell, R. R. Selvaraju, [et al.] // ArXiv. — 2016. — abs/1610.02424.

38. Findings of the Conf. on Machine Translation (WMT19) [Text] / L. Barrault, O. Bojar, M. R. Costa-jussà, [et al.] // Proc. Fourth Conf. on Machine Translation. V.2: Shared Task Papers. — Florence, 2019.

39. Findings of the WMT24 General Machine Translation Shared Task: The LLM Era Is Here but MT Is Not Solved Yet [Текст] / T. Kocmi [и др.] // Proceedings of the Ninth Conference on Machine Translation / под ред. B. Haddow [и др.]. — Miami, Florida, USA, 2024. — Нояб. — С. 1—46.

40. *Tessema, B. M.* UnifiedCrawl: Aggregated Common Crawl for Affordable Adaptation of LLMs on Low-Resource Languages [Text] / B. M. Tessema, A. Kedia, T.-S. Chung. — 2024. — arXiv: 2411.14343 [cs.CL].

41. P-Tuning: Prompt Tuning Can Be Comparable to Fine-tuning Across Scales and Tasks [Text] / X. Liu [et al.] // Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers) / ed. by S. Muresan, P. Nakov, A. Villavicencio. — Dublin, Ireland, 2022. — May. — P. 61—68.

42. A Natural Diet: Towards Improving Naturalness of Machine Translation Output [Text] / M. Freitag [et al.] // Findings of the Association for Computational Linguistics: ACL 2022 / ed. by S. Muresan, P. Nakov, A. Villavicencio. — Dublin, Ireland, 2022. — May. — P. 3340—3353.

43. A Family of Pretrained Transformer Language Models for Russian [Text] / D. Zmitrovich, A. Abramov, [et al.] // Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024). — Torino, Italia, 2024. — P. 507—524.

44. RuCoLA: Russian Corpus of Linguistic Acceptability [Text] / V. Mikhailov, T. Shamardina, [et al.] // Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. — Abu Dhabi, United Arab Emirates, 2022. — P. 5207—5227.

45. *Balikas*, *G.* On a Topic Model for Sentences [Text] / G. Balikas, M.-R. Amini, M. Clausel // Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval. — Pisa, Italy, 2016. — P. 921—924. — (SIGIR '16).

46. *Pitman*, *J.* The Two-Parameter Poisson-Dirichlet Distribution Derived from a Stable Subordinator [Text] / J. Pitman, M. Yor // Annals of Probability. — 1997. — Vol. 25. — P. 855—900.

47. Discourse Segmentation of Multi-Party Conversation [Text] / M. Galley [et al.] // ACL '03 Proceedings of the 41st Annual Meeting on Association for Computational Linguistics. — 2003. — Vol. 1. — P. 562—569.

48. *Hofmann*, *T.* Probabilistic Latent Semantic Indexing [Text] / T. Hofmann // Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. — New York, NY, USA, 1999. — P. 50—57.

49. *Robertson*, *S.* The Probabilistic Relevance Framework: BM25 and Beyond [Текст] / S. Robertson, H. Zaragoza // Foundations and Trends in Information Retrieval. — 2009. — Янв. — Т. 3. — С. 333—389.

50. *Prince*, *V.* Text Segmentation Based on Document Understanding for Information Retrieval [Текст] / V. Prince, A. Labadié //. — 06.2007.

51. *Dai*, *A. M.* Document Embedding with Paragraph Vectors [Текст] / A. M. Dai, C. Olah, Q. V. Le. — 2015. — arXiv: 1507.07998 [cs.CL].

52. Findings of the 2022 Conference on Machine Translation (WMT22) [Text] / T. Kocmi [et al.] // Proceedings of the Seventh Conference on Machine Translation (WMT) / ed. by P. Koehn [et al.]. — Abu Dhabi, United Arab Emirates (Hybrid), 2022. — Dec. — P. 1—45.

53. Qwen2.5 Technical Report [Text] / Qwen [et al.]. — 2025. — arXiv: 2412.15115 [cs.CL].

Публикации автора по теме диссертации

54. *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. and Skachkov, N. A. Hypotheses Re-ranking in Translation Models Using Human Markup // Teoria i sistemy upravlenia. 2024. No. 4. PP. 121-128.

55. *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.

56. *Skachkov, N.* Method of Input Masking for Training Translation Models [Text] / N. Skachkov // Pattern Recognition and Image Analysis. — 2025. — Vol. 35. — P. 493—500.

57. *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.

58. 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).

59. 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.