Федеральное государственное бюджетное
образовательное учреждение высшего образования
«Комсомольский-на-Амуре государственный университет»

На правах рукописи

Алена Анатольевна Животова

МАТЕМАТИЧЕСКАЯ МОДЕЛЬ, АЛГОРИТМЫ И ПРОГРАММНЫЙ КОМПЛЕКС ДЛЯ ПОВЫШЕНИЯ КАЧЕСТВА МАШИННОГО ПЕРЕВОДА УЗКОСПЕЦИАЛЬНЫХ ТЕХНИЧЕСКИХ ТЕКСТОВ НА АНГЛИЙСКИЙ ЯЗЫК

Специальность 1.2.2. –Математическое моделирование, численные методы и комплексы программ

Диссертация на соискание учёной степени
кандидата технических наук

Научный руководитель:
кандидат технических наук, доцент,
Бердоносов Виктор Дмитриевич

Комсомольск-на-Амуре – 2024

ОГЛАВЛЕНИЕ

ВВЕДЕНИЕ

Перевод – рутинная необходимость во многих отраслях, включая науку, производство, медицину и т.д., и с ростом количества информации и скорости ее генерирования растет и потребность в повышении качества перевода наряду с сокращением затрат на него. Современные системы машинного перевода (МП) показывают высокие показатели качества, кардинально изменив к 2023 г. структуру рынка лингвистических услуг, вытесняя переводчиков в пользу пост-редакторов и корректоров МП. Интерес исследователей к теме машинного перевода также значительно возрос. Так, согласно данным базы Science Direct количество статей по ключевым словам «машинный перевод» (Machine Translation) и «качество машинного перевода» (Machine Translation Quality) в 2023 году выросло на 117% и 146% соответственно по сравнению с 2017 годом [1].

В переводе специфика предметной области имеет ключевое значение, ведь МП тем эффективнее, чем больше обучающих данных (корпусов) загружено в систему, однако для некоторых предметных областей собрать достаточный объем параллельных тренировочных данных не всегда возможно. Так, например, нефтегазовый сектор – один из ключевых для экономики нашей страны с большой долей участия иностранных компаний в проектах освоения месторождений и нефтегазопереработки. Качество перевода в данной области имеет критически важное значение для коммуникации и обмена технологиями. Для подобных предметных узкоспециальных областей, связанных с объектами критической инфраструктурой, собрать достаточный объем двуязычных тренировочных корпусов проблематично ввиду ограничений конфиденциальности данных и секретности разработок.

Несмотря на выдающиеся прорывы нейросетевых, гибридных и больших языковых моделей МП в области семантической точности и гладкости перевода, вопрос качества перевода системами МП нельзя назвать решенным. Результат работы МП – черновик, который пользователь должен оценить и доработать самостоятельно. При этом пользователь без знания языка перевода не имеет инструментов для того, чтобы влиять на результат или хотя бы оценить качество полученного перевода. Эту

проблему активно освещают зарубежные исследователи A. Lear, C. Quinci, C. Canfora, A. Ottman, D. Kenny, P. Sanchez-Gijon. Предоставляя пользователю средства обработки текста на языке, носителем которого он является, на любом из этапов перевода, можно повысить его качество. Зная ключевые параметры текста и их связь с предполагаемой оценкой качества, становится возможным предложить алгоритмы и инструменты автоматического и/или полуавтоматического редактирования текста с целью его оптимизации для повышения качества перевода на требуемый язык.

Значительный вклад в разработку теоретических и практических основ в области подготовки исходных текстов к переводу, предварительного редактирования и упрощения естественных языков для систем автоматической обработки текстов, в частности систем МП, внесли зарубежные авторы: V. Kumar, F. Azadi, M. Federico, V. Alabau – в области интерактивного перевода; V. Sereton, P. Bouillon, J. Gerlach, A. Taufik, Y. Liang, W. Han, A. G. Arenas, C. Shei, Y. Hiraoka, M. Yamada, R. Miyata, A. Fujita – в области разработки подходов к предредактированию; L. O'Brien, D. Folaron, W. Aziz, M. Toledo – в области контролируемых и упрощенных языков. Среди российских авторов и для русского языка данная тема освещена незначительно, однако известны работы А.В. Ниценко, И. В. Оборневой, А.Д. Дмитриевой, А. Н. Лапошиной и др. в области оценки восприятия текста и упрощения русскоязычных текстов в соответствии с квалификацией реципиента.

Теоретическая актуальность и значимость темы определяется недостаточным уровнем исследований, касающихся алгоритмов предредактирования русскоязычных текстов для повышения качества их машинного перевода на другие языки, в частности на английский язык. Современные системы не анализируют исходный текст с целью оценки сложности задачи перевода и оптимизации результата МП, который должен быть проверен и при необходимости доработан пользователем, не всегда обладающим достаточной для этого компетенцией.

Практическая актуальность и значимость темы исследования объясняется тем, что в условиях развития экономики контента, когда цикл генерирования и обновления информации сократился с месяцев до дней, и с учетом необходимости ее локализации в режиме реального времени, требуется оптимизация временных и

материальных затрат на непрерывный перевод больших массивов текстовых данных с сохранением качества перевода, особенно в узкоспециальных технических областях, для который в открытых источниках не достаточно тренировочных данных. Использование вероятностной оценки сложности задачи перевода и алгоритмов оптимизационного предредактирования исходных текстов позволяет снизить зависимость качества МП от человеческого фактора, а также предоставляет необходимые критерии для разработки стратегии управления рисками, связанными с компетенцией исполнителей и пользователей систем МП, при реализации крупных переводческих проектов.

Основная идея диссертации в том, чтобы, используя особенности работы алгоритмов систем МП и основы теории перевода, автоматизировать предварительное редактирование исходных текстов с тем, чтобы оптимизировать их структуру, благодаря чему системы МП будут эффективнее переводить их на требуемый язык и допускать меньше стилистических ошибок, для распознания которых требуется более высокая компетенция пользователя в области языка перевода.

Объектом исследования выступает процесс перевода текстов, **предметом** исследования – методы повышения качества перевода при работе с исходным текстом.

Целью работы является разработка моделей и алгоритмов и их реализация для повышения качества машинного перевода узкоспециальных технических текстов путем автоматического оптимизационного предредактирования.

Задачи исследования**:**

- Выполнить анализ существующих систем машинного перевода, направлений их совершенствования и способов реализации автоматического оптимизационного предредактирования.
- Разработать математическую модель процесса перевода.
- Разработать методику и алгоритм вероятностной оценки сложности задачи перевода.
- Разработать методику оптимизационного предредактирования исходных текстов.

- Реализовать разработанные алгоритмы в программном комплексе для оценки сложности задачи перевода русскоязычных текстов на английский язык и оптимизационного предредактирования с целью повышения качества перевода на английский язык.
- Проверить адекватность разработанных алгоритмов на корпусе узкоспециальных технических текстов.

Научная новизна:

1. Предложена новая методика для повышения качества машинного перевода текстов с русского языка на английский язык, отличающаяся от существующих применением обратного перевода для сбора тренировочных данных и оптимизационного предредактирования на основе вероятностной оценки сложности задачи перевода.

2. Впервые предложена методика оценки сложности переводческой задачи для переводчика на основе его компетенции и специализации и параметров исходного текста, которая позволяет прогнозировать риски некачественного и/или несвоевременного решения задачи перевода.

3. Предложен новый алгоритм, позволяющий расширить область применения оптимизационного метода градиентного спуска путём использования элементов нечеткой логики в выражении функции правдоподобия через функцию принадлежности полученного текста низкой сложности задачи перевода для выбранной системы МП.

4. Предложен новый алгоритм, позволяющий расширить область применения метода наименьших квадратов для поиска весов значимости параметров исходного текста для вероятностной оценки ожидаемого качества его перевода на целевой язык.

5. Предложена новая архитектура и реализован программный комплекс для повышения качества машинного перевода текстов с русского языка на английский язык, отличающийся от существующих применением ансамбля моделей для оптимизационного предредактирования на основе вероятностной оценки сложности задачи перевода с целью повышения качества машинного перевода текстов с русского языка на английский язык.

Основные положения, выносимые на защиту:

1. Математическая модель процесса перевода, позволяющая определить понятия компетенции и специализации переводчика через отношения множеств.

2. Математическая постановка задач машинного перевода и оптимизационного редактирования при помощи функции правдоподобия и элементов нечеткой логики.

3. Решение задачи максимизации правдоподобия оптимизационного предредактирования численным методом градиентного спуска (подъема).

4. Методика и алгоритм вероятностной оценки сложности задачи перевода на основе регрессионного анализа зависимости ожидаемого качества машинного перевода от признаков русскоязычного текста с оптимизацией функции потерь методом наименьших квадратов.

5. Методика и алгоритмы оптимизационного предредактирования русскоязычных текстов, позволяющие сократить объем необходимых данных для тренировки модели путем применения концепции обратного перевода и время тренировки модели за счет предварительной обработки тренировочных данных.

6. Программный комплекс для реализации системы анализа и предредактирования русскоязычных текстов для повышения качества машинного перевода на английский язык.

Практическая значимость работы обусловлена возможностью интегрирования программного обеспечения, реализующего вышеперечисленные алгоритмы, основанные на вероятностной оценке сложности задачи перевода и алгоритмах оптимизационного редактирования, в системы управления и автоматизации переводческой деятельности. Разработанные алгоритмы позволят повысить качество перевода русскоязычных узкоспециальных технических текстов в условиях ограниченного объема эталонных двуязычных корпусов для обучения нейросетевых моделей, оптимизировать затраты на перевод, повысить надежность существующих систем, снизить зависимость качества перевода от человеческого фактора.

Разработанные в ходе исследования алгоритмы и программные комплексы реализованы и внедрены в практическую деятельность ведущего предприятия

лингвистической отрасли в г. Комсомольске-на-Амуре – ООО «Агентство переводов «ФИАС-Амур» (акт внедрения результатов диссертации на соискание ученой степени кандидата технических наук № 6/23/1 от 10.06.2023); получены свидетельства о регистрации программ ЭВМ и БД для 6 программных модулей, 2 программных комплекса, 2 баз данных.

Достоверность результатов исследования определяется применением апробированных математических методов, включая теорию множеств, численных методов оптимизации, таких как метод наименьших квадратов и метод градиентного спуска, статистических методов, а именно метода максимального правдоподобия, а также использованием современных комплексов программ анализа данных и экспериментально.

Личный вклад автора. Все результаты, представленные в работе, получены автором самостоятельно. Из совместных работ в работу включены только результаты, полученные лично автором. Соавторы публикаций по теме диссертации участвовали в обсуждении постановочной части решаемых задач и результатов, полученных по разработанным автором методам и алгоритмам.

Соответствие паспорту специальности. Диссертационная работа со-ответствует области исследования специальности 1.2.2 «Математическое моделирование, численные методы и комплексы программ» по п. 2 «Разработка, обоснование и тестирование эффективных вычислительных методов с применением современных компьютерных технологий» (п. 1,2 научной новизны), п. 3 «Реализация эффективных численных методов и алгоритмов в виде комплексов проблемно-ориентированных программ для проведения вычислительного эксперимента» (п. 5 научной новизны), п. 4 «Разработка новых математических методов и алгоритмов интерпретации натурного эксперимента на основе его математической модели» (п. 3,4 научной новизны), п. 8 «Комплексные исследования научных и технических проблем с применением современной технологии математического моделирования и вычислительного эксперимента» (п. 1-4 научной новизны).

Апробация результатов исследования. Основные результаты работы докладывались и обсуждались на следующих научных конференциях:

– краевой конкурс молодых ученых Хабаровского края «XXVI краевой конкурс молодых ученых в сфере научных исследований», I место в секции «Физико-математические науки и информационные технологии» (г. Хабаровск, 2024 г.);

– 29-ая международная конференция по компьютерной лингвистике и интеллектуальным технологиям «Диалог» (г. Москва, 2023 г.);

– международная научно-практическая конференция «Информационные технологии и интеллектуальные системы принятия решений» (ITIDMS 2022) (г. Москва, 2023 г.);

– VII международная научно-практическая конференция «Информационные технологии и высокопроизводительные вычисления» (ITHPC-2023), диплом III степени за лучший оклад среди молодых ученых (г. Хабаровск, 2023 г.);

– краевой конкурс молодых ученых Хабаровского края «XXV краевой конкурс молодых ученых в сфере научных исследований» (г. Хабаровск, 2023 г.);

– VI всероссийская национальная научная конференция молодых учёных «Молодёжь и наука: актуальные проблемы фундаментальных и прикладных исследований» (г. Комсомольск-на-Амуре, 2023 г.);

– международный конкурс «2022 International Activity of Innovation Entrepreneurship Creation», организован представительством МА ТРИЗ в Китае при поддержке Китайской ассоциации по науке и технологиям (Китай, 2022 г.);

– II международная научно-практическая конференция молодых ученых «Наука, инновации и технологии: от идей к внедрению» (г. Комсомольск-на-Амуре, 2022 г.);

– международная мультидисциплинарная конференция по промышленному инжинирингу и современным технологиям «Far East Con-2020» Дальневосточного федерального университета (г. Владивосток, 2020 г.).

Публикации. Основные теоретические и практические результаты диссертационного исследования опубликованы в 12 научных работах, в том числе в 3 работах в издании, рекомендованном ВАК, в 3 работах в изданиях, индексируемых в международной базе Scopus.

Объём и структура работы. Диссертация включает в себя введение, четыре основные главы, заключение, список используемой литературы и 2 приложения, изложена на 125 страницах. Текст работы содержит 13 таблиц и 22 рисунка и список литературы из 124 наименований. В приложениях содержатся копии свидетельств о государственной регистрации программ для ЭВМ и баз данных, копия акта внед-рения результатов диссертации.

В первой главе предложено использовать ТРИЗ-эволюционный подход к выявлению направлений развития гибридных систем нейронного МП, сформулированы положения ТРИЗ-эволюционного анализа, позволяющие систематизировать исследуемую область знаний с высокой степенью детализации. Применение ТРИЗ-эволюционного анализа систем МП позволило: систематизировать данные об эволюции систем МП; определить ключевые этапы развития систем МП; выделить главные производственные параметры, определяющие направления развития систем нейронного МП. Показано, что перспективным направлением является развитие методов и алгоритмов автоматизированной предобработки исходных текстов для нейронного МП, при этом исследование в этой области позволит добиться повышения качества МП. ТРИЗ-эволюционная карта позволила определить проблемы повышения качества МП, которые могут быть решены путем оптимизационного предредактирования исходных текстов. Кроме того, в главе рассмотрены основные способы предобработки исходных текстов для систем МП, таких как использование контролируемого языка, правил предредактирования и решение задачи перефразирования текста в контексте решаемой задачи.

Во второй главе предложено решение задачи оптимизационного предредактирования методом градиентного спуска (подъема); приводится обобщенная модель процесса перевода, разработанная на основе теории множеств, описывающая систему понятий прикладной лингвистики в области машинного перевода, включая формализацию понятий опыта, специализации и компетенции переводчика; описаны математическая постановка задач перевода и оптимизационного предредактирования через функцию правдоподобия и с использованием элементов нечеткой логики. В ходе моделирования впервые обоснована целесообразность и разработана

методология оценки сложности переводческой задачи. Результаты выполненного моделирования показывают, что уже на этапе оценки исходного текста, возможно предсказать ожидаемое качество перевода на основе параметров исходного текста и компетенции и специализации переводчика.

В третьей главе рассматриваются методика осуществления автоматического оптимизационного предредактирования текста с целью повышения качества машинного перевода относительно формализованных требований и методика расчёта сложности задачи перевода заданного текста для заданного переводчика в соответствии с формализованными требованиями к переводу.

В четвёртой главе описана реализация методов автоматического оптимизационного предредактирования узкоспециального технического русскоязычного текста с целью повышения качества его МП на английский язык и оценки сложности задачи перевода для системы МП; описывается программный комплекс, реализованный в соответствии с этими методиками. Описывается архитектура программного комплекса и основных его подсистем: генератора МП текстов с русского языка на английский язык; препроцессинга текстовых данных для взвешенной оценки параметров русскоязычного текста; вероятностной оценки сложности переводческой задачи для систем МП; автоматической очистки сырых данных из памятей переводов CAT для тренировки языковой модели; тренировочного модуля языковой модели для перефразирования русскоязычных технических текстов, предредактора русскоязычных узкоспециальных текстов для систем МП. Описаны полученные в ходе реализации описанных методов и алгоритмов массивы данных: база данных показателей структурного анализа предложений технических русскоязычных текстов, корпус параллельных двуязычных текстов нефтегазовой тематики для тренировки языковых моделей в задачах перефразирования узкоспециальных технических русскоязычных текстов и повышения качества их перевода на английский язык. Приводятся результаты тестирования программного корпуса на реальных данных. Описаны возможности интеграции программного комплекса в контур автоматизации процессов переводческой деятельности.

ГЛАВА 1 ПРЕДРЕДАКТИРОВАНИЕ КАК ПЕРСПЕКТИВНОЕ НАПРАВЛЕНИЕ РАЗВИТИЯ СИСТЕМ МАШИННОГО ПЕРЕВОДА

1.1 Развитие систем машинного перевода

Одним из способов решения задачи повышении качества перевода наряду с сокращением затрат на него является автоматизация процессов перевода. Первые попытки такой автоматизации берут свое начало еще в 1950х годах. С тех пор технологии МП прошли большой путь, но только в 2014-2017 гг. [2, 3, 4] произошел качественный скачок роста идеальности данных систем, который привел рынок лингвистических услуг к пониманию того, что внедрение и развитие данных технологий является одним из наиболее перспективных направлений развития отрасли [5, 6]. Современные системы МП показывают высокие показатели качества, кардинально изменив к 2023 г. структуру рынка лингвистических услуг России [7], вытесняя переводчиков в пользу пост-редакторов и корректоров машинного перевода.

В литературе уже описаны многие аспекты истории развития систем и технологий МП. Автором одних из наиболее подробных обзоров является John Hutchins [8, 9, 10], в которых подробно описаны ключевые этапы развития систем на момент издания, их достоинства и недостатки и т.д. Многие исследователи описывают историю развития МП в виде хронологического обзора с целью дать представление об истории развития предметной области [11, 12, 13] описывают в большей степени социальный аспект развития технологий МП. Также в литературе приведены различные варианты классификации технологий МП по используемым методам, по структуре используемых для обучения данных и тп. Так, например, Lane Schwartz [14] очень подробно описывает историю развития МП, основные парадигмы и классификации технологий со ссылками на авторов и разработчиков, а также приводит обзор литературы по данной теме. Из всех работ по классификации технологий МП и категоризации статей по данной теме отдельно следует отметить архив статей по МП [15] – картотека материалов по МП за период с 1950 по 2017 год. В ней не только

собрана ключевая информация по МП, но и проведена подробнейшая категоризация по технологиям и разделам данной предметной области.

Несмотря на широкую освещенность темы в литературе, многие из существующих статей, опубликованных в рецензируемых журналах, устарели. Технологии за последние годы совершили качественный скачек, но рецензируемых материалов, описывающих новые технологии в разрезе эволюции очень мало, а описание перспектив и направлений развития ограничено и/или отсутствует их обоснование. Систематизация информацию о технологиях МП необходима выявления ключевых направлений и перспектив развития и исследования. Предлагается в анализ и систематизацию включить: краткое изложение ключевых аспектов технологии МП; оценку развития главных параметров при переходе от одной технологии к другой; определение ключевых проблем, ограничивающих применение в каждой технологии; компактное, но емкое описание решений, за счет которых произошло преодоление выявленных ограничений.

Такой анализ позволит получить представление о развитии систем МП, составить карту проблем, требующих решения, что сократит время на выявление актуальных проблем МП за счёт систематизации, визуализации и структурирования ключевых данных, а определение релевантных путей совершенствования систем МП облегчит и ускорит постановку задач исследований и формулирования гипотез.

1.1.1 Анализ на основе ТРИЗ-эволюционного подхода

Развернувшаяся в последние десятилетия информационная революция обострила и потребности создания и развития новых методов извлечения и систематизации знаний [16], которые обеспечивали бы возможность изучения и структурирования огромного объёма информации за ограниченное время. Наметить пути разрешения данного противоречия позволяет ТРИЗ-эволюционный подход. ТРИЗ (теория решения изобретательских задач) – область знаний, исследующая механизмы развития искусственных систем с целью создания практических методов решения инновационных задач [17]. Благодаря своей высокой эффективности и универсальности, ТРИЗ получила международное признание, успешно применяется и развивается во

многих областях деятельности человека, прежде всего, в промышленном производстве, науке и образовании [18, 19].

Согласно ТРИЗ, в развитии искусственных систем происходит чередование этапов количественного роста и качественных скачков. В процессе количественного роста в результате неравномерного развития характеристик искусственной системы проявляются противоречия, которые препятствуют повышению идеальности системы. Идеальность, в ТРИЗовском понимании, это отношение суммы параметров, характеризующих пользу к сумме параметров, характеризующих затраты. Как правило, в качестве оценки идеальности выступает главный производственный параметр. Противоречие – это проявление несоответствия между требованиями, предъявляемыми к системе, и ограничениями, налагаемыми на нее. Выявление и анализ противоречий лежат в основе прогнозирования развития систем. Разрешение противоречий неизменно ведет к повышению идеальности системы.

ТРИЗ-эволюционный подход [20] предполагает выстраивать эволюцию исследуемых систем по мере увеличения их идеальности. Развитие системы происходит при разрешении противоречий, за счёт использования инструментов ТРИЗ [21] (приёмы разрешения противоречий, вепольный анализ, законы и тренды развития систем и так далее). Таким образом, выстраивается эволюция систем от противоречия к противоречию. Такое выстраивание, позволяет не только систематизировать знания по соответствующим системам, но и предлагать новые высокоэффективные решения. Анализ при помощи ТРИЗ-эволюционного подхода позволяет систематизировать знания по соответствующим системам; определить главные производственные параметры, выявить закономерности и перспективы развития с учетом повышения идеальности; предлагать новые высокоэффективные решения путём разрешения выявленных противоречий. Данный подход был многократно описан в литературе на примере таких предметных областей как объектно-ориентированное программирование [22], автоматизированные системы управления [23], установки коксования [24] и др.

Согласно ТРИЗ-эволюционному подходу [25] исследование развития некоторой системы включает три этапа: анализ развития системы (I), построение ТРИЗ-

эволюционной карты (II) и анализ ТРИЗ-эволюционной карты (III). Этап I включает следующие шаги:

Шаг 1. Описание исходного объекта ТРИЗ-эволюции.

Шаг 2. Выявление противоречий у выбранного объекта. Противоречия записываются по формуле:

$$Con_i{:}MP_j{,}MP_k, \tag{1}$$

где MP_j – улучшаемый параметр, MP_k – ухудшаемый параметр.

Шаг 3. Определение инструментов ТРИЗ, позволяющих разрешить выявленные противоречия.

Шаг 4. Описание итераций ТРИЗ-эволюции, т.е. переходов к последующим объектам, в которых разрешены отдельные противоречия. Описание содержит краткое текстовое описание разрешение противоречия(ий) и символьное описание итерации по формуле:

$$\{Con_i,\ldots,Con_n\} \xrightarrow{(\{IP_j,\ldots,\ IP_m\})} Sol_x{:}\{MP_k,\ \ldots,\ MP_y\}, \tag{2}$$

где $Con_i,\ldots,Con_n$ – противоречия, разрешенные в рамках итерации ТРИЗ-эволюции; $IP_j,\ldots,\ IP_m$ – инструменты ТРИЗ, позволяющие сформировать и/или описать решение при переходе от одного объекта ТРИЗ-эволюции к другому; Sol_x – решение, согласно которому разрешены перечисленные противоречия; $MP_k,\ \ldots,\ MP_y$ – изменяемые в результате решения параметры.

Далее *Шаги 1-4* повторяются для всех наиболее значимых объектов исследуемой области.

На этапе II происходит графическое представление итераций ТРИЗ-эволюции выявленных и описанных на этапе I. Графическая форма итерации представлена на рисунке 1.1.

На этапе III ТРИЗ-эволюционная карта анализируется с целью определить неразрешенные противоречия и противоречия, разрешение которых привело к повышению идеальности не за счет устранения мешающего параметра. Данные противоречия позволят сформулировать постановку задачи перспективных направлений развития системы. Кроме того, по ТРИЗ-эволюционной карте можно отследить качественные скачки роста идеальности и ключевые тенденции развития системы.

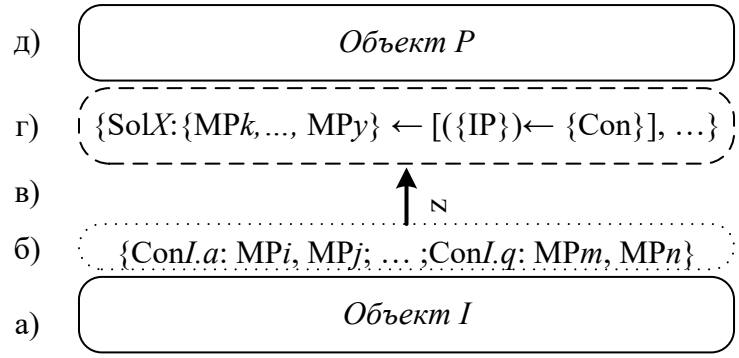

Рисунок 1.1 – Графическое представление итерации ТРИЗ-эволюционной карты: (а) – исходный объект итерации; (б) – множество противоречий исходного объекта итерации; (в) – переход к конечному объекту итерации, где Z – порядковый номер итерации на ТРИЗ-эволюционной карте; (г) – множество разрешений противоречий исходного объекта итерации, реализованных в конечном объекте итерации; (д) – конечный объект итерации

1.1.2 ТРИЗ-эволюционный анализ систем МП

Отправной точкой автоматизации МП можно считать 1933 год [26]. Именно в этом году впервые были запатентованы независимо друг от друга две машины для осуществления перевода: «Механический мозг» Жоржа Арцруни и «Машина для подбора и печатания слов при переводе с одного языка на другой или на несколько других одновременно» П. П. Троянского (патент СССР № 40995 от 5 сентября 1933 г.) [27]. Развиваться же направление МП стало гораздо позднее, во времена холодной войны.

Первым подходом к автоматизации перевода считается пословный перевод [28, 29, 30]. Ключевая идея: разделяем предложение по словам, находим перевод по двуязычному словарю, подставляем перевод на место каждого слова. Недостатки подхода:

- чем сложнее предложение, тем хуже качество перевода;
- требует долгого и кропотливого формирования словарей, что трудозатратно и ограничивает количество языковых пар;
- не учитывает разницу грамматический строй, морфологию, согласование падежей;
- не учитывает контекстно-зависимые значения слов.

Исходя из описанных недостатков опишем противоречия данного подхода, используя формулу (1) (см. Таблицу 1.1).

Таблица 1.1 – Спецификация противоречий пословного подхода к переводу

Описание противоречия	**Формула противоречия**
Противоречие 1.1: при увеличении количества возможных смысловых значений одной лексической единицы недопустимо снижается качество перевода	Con1.1: MP2↑, MP1.1↓
Противоречие 1.2: при повышении качества перевода работоспособность системы недопустимо снижается	Con1.2: MP1↑, MP3↓
Противоречие 1.3: при увеличении количества языковых пар недопустимо увеличивается время на составление обучающих данных	Con1.3: MP4↑, MP5↑
Противоречие 1.4: при повышении сложности структуры входных данных качество перевода недопустимо снижается	Con1.4: MP6↑, MP1↓

Рассмотрим пример разрешения противоречия Con1.2. Противоречие может быть разрешено при помощи приема «Принцип предварительного действия» (IP10): предварительно сформировать набор правил, в соответствии с которыми будет обрабатываться текст. Данное решение было реализовано в системах дословного перевода. Произошли первая итерация ТРИЗ-эволюции и переход к новой парадигме «Машинный перевод на основе правил». Представим данную итерацию в символьном виде согласно формуле (2):

$$Con1.2: MP1\uparrow, MP3\downarrow \xrightarrow{IP:10} Sol1:MP3\uparrow$$

Графически первая итерация ТРИЗ-эволюции систем МП представлена на рисунке 1.2. Полный перечень итераций ТРИЗ-эволюции от исходного объекта приведен в таблице 1.2. В качестве объектов ТРИЗ-эволюции можно выделить следующие системы МП: дословный МП [31], трансферный МП [32], интерлингвистический МП [33, 34], МП на примерах [35], интерактивный МП [36], статистический МП по словам [37], статистический МП по фразам [38], статистический МП на основе синтаксиса [39], нейронный МП [40], нейронный МП «без учителя» [41], адаптированный нейронный МП [42], гибридный МП [43]. Далее рассмотрим краткую характеристику каждого вида указанных систем МП.

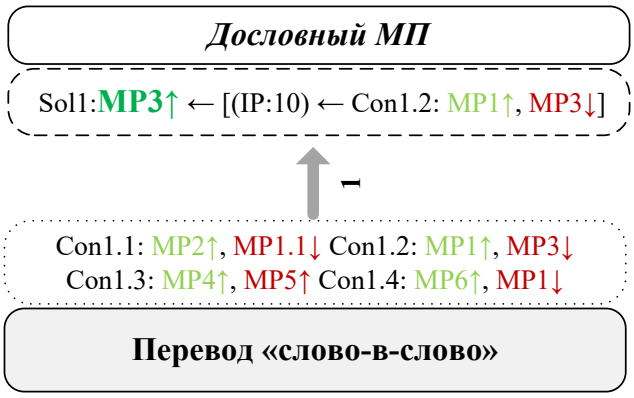

Рисунок 1.2 – Первая итерация ТРИЗ-эволюции систем машинного перевода

Дословный МП: текст разделяется по словам; система ищет перевод каждого слова на основе правил; на основе правил правятся ошибки морфологии, согласования падежей, использования окончаний и т.д.

Трансферный МП: текст на исходном языке анализируется для определения его грамматической структуры; грамматическая структура (члены предложения) перестраивается в структуру, подходящую для воспроизводства текста на языке перевода; выполняется подстановка дословного перевода каждой отдельной части текста в нужное место конечной структуры).

Интерлингвистический МП: исходное предложение при помощи словаря и грамматических правил исходного языка полностью конвертируется в промежуточное представление (совокупность концептов), единое для всех языков мира (interlingua), с последующей конвертацией его в текст на языке перевода при помощи словаря и грамматических правил языка перевода.

МП на примерах: в систему загружаются базы переводов, которыми являются двуязычные корпусы текста (оригинал-перевод); система выполняет поиск ранее переведенных частей, определяет отличные части текста и переводит только отличные от хранящихся в базе переводов части, подставив остальные части из готового перевода.

Интерактивный МП: на разных стадиях перевода привлекается участие человека. Такое участие может быть выражено в разных формах: с постредактированием, когда человек редактирует уже переведенный машиной текст; с

предредактированием, когда человек редактирует входной текст, приспосабливая его для более легкого понимания машиной; частично автоматизированный перевод, когда человек и машина взаимодействуют в процессе перевода, и участие человека в процессе необходимо для разрешения трудностей.

Статистический МП по словам: система обучается на параллельных корпусах которые разделены на слова и проанализированы по критерию частотности отдельных слов; система определяет порядок слов в переведенных предложениях, в том числе относительный порядок; выполняется перевод отдельных слов путем подстановки статистически более вероятных вариантов на основе обучающих данных; после перевода слова меняются местами согласно статистически более естественному положению; на те места, где предполагается необходимость нового слова, вставляются маркеры (NULL-слова); под каждый маркер выполняется подбор нужного слова (артикля, частицы или глагола).

Статистический МП по фразам: принцип работы совпадает с системами статистического МП по словам с той разницей, для обучения и перевода текст делится не только на слова, но и на n-граммы – пересекающиеся наборы из N слов подряд.

Статистический МП на основе синтаксиса: перед переводом выполняется точный синтаксический разбор предложения, на основании чего строится дерево, используя которое, теоретически, можно обучить систему правильно конвертировать грамматические структуры одного языка в структуры другого, выполняя перевод по словам или фразам. Практически данный подход не был реализован до конца из-за сложности задачи автоматического синтаксического разбора.

Нейронный МП: система состоит из двух нейронных сетей, каждая из который знает только один язык; одна сеть кодирует предложение на языке оригинала в набор цифровых характеристик, а вторая, проанализировав предложение до конца, причем одновременно и слева направо, и справа налево, начинает декодировать числовые характеристики и предсказывать слова перевода, причем каждое предсказанное слово используется для предсказания следующего слова; для более точного выбора слов для перевода используется контекст всего исходного предложения, а также

контекст всех предыдущих предсказанных слов; система обучается на двуязычных корпусах посредством алгоритмов глубокого обучения.

Адаптированный нейронный МП: система нейронного перевода, корпуса для обучения которой составляют тексты некоторой предметной области, исключая попадания в обучающий корпус текстов других предметных областей/тематик.

Нейронный МП «без учителя»: система строится на нейронных сетях, архитектура которых позволяет производить обучение «без учителя» на моно-язычных корпусах текстов, то есть без эталонных данных для проверки перевода.

Гибридный МП: система, в которую интегрированы разные технологии МП для достижения лучшего качества перевода. Наиболее часто в таких системах интегрируют нейронный и статистический МП.

В таблице 1.2 представлен перечень выявленных в рамках анализа главных параметров систем МП.

Таблица 1.2 – Описание главных параметров систем МП

Обозначение	Описание
МР1	Качество перевода (МР1.1 – с точки зрения лексической точности; МР1.2 – с точки зрения грамматической точности; МР1.3 – с точки зрения стилистической точности; МР1.4 – с точки зрения единообразия; МР1.5 – с точки зрения смысловой точности; МР1.6 – гладкость перевода; МР1.7 – длинных предложений; МР1.8 – коротких предложений; МР1.9 – текстов узких тематик)
МР2	Количество возможных смысловых значений одной лексической единицы
МР3	Работоспособность системы
МР4	Количество языковых пар
МР5	Время на подготовку обучающих данных (МР5.1 – двуязычных словарей; МР5.2 – параллельных корпусов; МР5.3 – правил; МР5.4 – моно-корпусов)
МР6	Сложность структуры текста оригинала
МР7	Объем обучающих данных (МР7.1 – правил; МР7.2 – корпусов текста)
МР8	Время на разработку системы
МР9	Трудозатраты на сопровождение системы
МР10	Количество вариантов сочетаний слов
МР11	Трудозатраты на пред-, пост-редактирование
МР12	Время на обучение системы
МР13	Вероятность грубых ошибок
МР14	Объем исходного текста
МР15	Время на поиск ошибок в переведенном тексте
МР16	Количество переводимых тематик
МР17	Сложность системы

Для каждой из систем произведен подробный анализ с выявлением ключевых проблем, ограничивающих рост идеальности, на основании которых сформулированы противоречия. Спецификация выявленных проблем и противоречий в системах МП представлена в таблице 1.3.

Таблица 1.3 – Спецификация противоречий систем машинного перевода

Описание проблемы	Противоречие	Символьное представление
Дословный МП		
Практически невозможно язык представить в виде лишь набора правил, следовало бы также учесть и все возможные исключения из правил (неправильные глаголы в английском, плавающие приставки в немецком, суффиксы, диалекты, сленг и т.д.	Противоречие 2.1: при повышении качества перевода недопустимо увеличивается объем обучающих данных (правил)	Con2.1: MP1↑, MP7.1↑
Количество правил в каждом языке огромно и для качественной проработки необходимых обучающих данных необходимо огромное количество человеко-часов	Противоречие 2.2: при повышении качества перевода недопустимо увеличивается время на подготовку обучающих данных	Con2.2: MP1↑, MP5↑
Необходимо постоянно поддерживать лингвистическую базу в актуальном состоянии, так как язык - динамическая система	Противоречие 2.3: при повышении качества перевода недопустимо увеличиваются трудозатраты на сопровождение системы	Con2.3: MP1↑, MP9↑
Трансферный МП		
С одной стороны можно задать общие правила переноса грамматической структуры, что упрощает задачу перевода, с другой стороны, сочетаний слов намного больше, чем самих слов, и каждый вариант почти невозможно учесть.	Противоречие 3.1: при сокращении времени на разработку системы недопустимо увеличивается объем обучающих данных (правил)	Con3.1: MP9↓, MP7.1↑
	Противоречие 3.2: при сокращении количества правил недопустимо снижается качество перевода возможных сочетаний слов	Con3.2: MP7.1↓, MP1.1↓
Интерлингвистический МП		
Сложность реализации и отсутствие методов и моделей поиска закономерностей и классификации атрибутов текста для создания унифицированного языка и его структуры.	Противоречие 4.1: при повышении качества перевода недопустимо увеличивается время на разработку системы	Con4.1: MP1↑, MP8↑
МП на примерах		
Примеры содержат слова, словосочетания и даже предложения, но, фактически, мы находим дословно схожие части, не учитывая особенности синтаксиса, морфологии, грамматического	Противоречие 5.1: при повышении качества перевода объем обучающих данных (корпусов текста) недопустимо увеличивается	Con5.1: MP1↑, MP7.2↑

Описание проблемы	Противоречие	Символьное представление
строя и т.д. Чтобы учесть все возможные варианты, необходимо больше обучающих данных.		
	Противоречие 5.2: при увеличении количества вариантов сочетаний слов в тексте оригинала качество перевода с точки зрения единообразия недопустимо снижается	Con5.2: MP10↑, MP1.4↓
Даже если объем обучающих данных достаточный, система не делит предложение на структурные части из-за чего, например, служебные части речи, влияющие на контекст, могут отразится на качестве перевода.	Противоречие 5.3: при повышении качества с точки зрения единообразия недопустимо снижается качество с точки зрения передачи смысла	Con5.3: MP1.4↑, MP1.5↓
Из-за разницы структуры языков перевода опущенные (нулевые) части предложений не учитываются в переводе, либо переводятся части предложения, которые ввиду грамматических правил языка перевода должны быть опущены.	Противоречие 5.4: при повышении качества с точки зрения единообразия недопустимо снижается качество перевода с точки зрения грамматической точности	Con5.4: MP1.5↑, MP1.2↓
Интерактивный МП		
Система не может работать автоматически без участия человека	Противоречие 6.1: при повышении качества перевода недопустимо увеличиваются трудозатраты на пред-, пост-редактирование	Con6.1: MP1↑, MP11↑
Статистический МП по словам		
При подготовке обучающих данных необходимо максимально точное соответствие оригинала и перевода. Однако, не всегда перевод может быть строго формализован, есть еще литературные или вольные переводы, которые также необходимо учитывать.	Противоречие 7.1: при повышении качества перевода недопустимо увеличивается время на подготовку обучающих данных (корпусов)	Con7.1: MP1↑, MP5.2↑
Из-за отсутствия двуязычных словарей между некоторыми языками, система переводит текст сначала на английский, а затем на язык перевода из-за чего возникают «двойные потери» качества.	Противоречие 7.2: при повышении количества языковых пар недопустимо снижается качества перевода	Con7.2: MP4↑, MP1↓
Статистический МП по фразам		
Статистические аномалии	Противоречие 8.1: при увеличении объема обучающих данных недопустимо снижается качество перевода с точки зрения смысловой точности	Con8.1: MP7↑, MP1.5↓
Отдельные фразы плохо согласуются между собой, в итоге переведенное предложение – набор фраз, иногда не связных по смыслу	Противоречие 8.2: при повышении качества с точки зрения лексической точности недопустимо снижается гладкость перевода	Con8.2: MP1.1↑, MP1.6↓
	Противоречие 8.3: при повышении качества с точки зрения лексической точности недопустимо снижается смысловая точность	Con8.3: MP1.1↑, MP1.5↓

Описание проблемы	Противоречие	Символьное представление
Статистический МП на основе синтаксиса		
Даже для корпуса с простейшими 2-3 уровневыми деревьями, время обучения слишком велико, а значит на практике система не может быть применима.	Противоречие 9.1: при повышении качества перевода с точки зрения грамматической точности время на обучение системы недопустимо увеличивается	Con9.1: MP1.2↑, MP12↑
Не для всех языков разработаны методы синтаксического разбора. Не для всех языков, для которых разработаны методы синтаксического анализа, они работают достаточно качественно.	Противоречие 9.2: при повышении количества языковых пар недопустимо снижается качество перевода	Con9.2: MP4↑, MP1↓
Нейронный МП		
В целом «гладкий» перевод может содержать грубые лексические ошибки.	Противоречие 10.1: при повышении гладкости перевода недопустимо повышается вероятность наличия в переводе грубых ошибок	Con10.1: MP1.6↑, MP13↑
Почти любой перевод требует понимания контекста нескольких предложений, иногда это имеет решающее значение для точного перевода с точки зрения используемой лексики. Нейронная система не может анализировать и хранить информацию о контексте текста большого объема и эффективно ее запоминать.	Противоречие 10.2: при увеличении объема исходного текста недопустимо снижается качество перевода с точки зрения смысловой точности	Con10.2: MP14↑, MP1.5↓
Зависимость от состава обучающих данных: если тренировать нейронный перевод только на длинных парах предложений, система будет неспособна перевести корректно короткое предложение или даже отдельное слово.	Противоречие 10.3: при повышении качества перевода длинных предложений недопустимо снижается качество перевода коротких предложений	Con10.3: MP1.7↑, MP1.8↓
Аномалии в переводе: пропущенные отрицания, отдельные слова или целые фразы. Аномалии непредсказуемы и непоследовательны, что затрудняет их автоматическое выявление и исправление.	Противоречие 10.4: при повышении гладкости перевода недопустимо увеличивается время на поиск возможных ошибок	Con10.4: MP1.6↑, MP15↑
Низкое качество перевода исходных текстов, которые сильно отличаются от данных, использованных для машинного обучения.	Противоречие 10.5: при увеличении количества тематик исходного текста недопустимо увеличивается время на подготовку обучающих данных	Con10.5: MP16↑, MP5↑
	Противоречие 10.6: при увеличении количества тематик исходного текста недопустимо снижается качество перевода	Con10.6: MP16↑, MP1.9↓

Описание проблемы	Противоречие	Символьное представление
Адаптированный нейронный МП		
По узким тематикам чрезвычайно сложно производить сбор и обработку релевантных двуязычных корпусов для обучения нейронной сети в достаточном объеме.	Противоречие 11.1: при повышении качества перевода текстов узкой тематики и/или редкой языковой пары недопустимо увеличивается время на подготовку обучающих данных	Con11.1: МР1.9↑, МР5.2↑
Нейронный МП «без учителя»		
Необходимость подготовки моно-корпусов большого объема по редким языкам и/или тематикам в векторном представлении.	Противоречие 12.1: при повышении качества перевода редких языковых пар или тематик объем необходимых обучающих данных недопустимо увеличивается	Con12.1: МР1↑, МР7↑
Гибридный МП		
При комбинировании разных систем достигается более высокое качество перевода, но это означает усложнение системы, которая наследует не только преимущества, но и недостатки систем, входящих в ее состав.	Противоречие 13.1: при повышении качества перевода недопустимо увеличивается сложность конечной системы	Con13.1: МР1↑, МР17↑
	Противоречие 13.2: при увеличении количества тематик недопустимо увеличивается время на разработку системы	Con13.2: МР16↑, МР8↑
	Противоречие 13.3: при увеличении количества языковых пар недопустимо увеличивается время на разработку системы	Con13.3: МР4↑, МР8↑

В таблице 1.4 произведен анализ выявленных противоречий и инструментов ТРИЗ, при помощи которых были разрешены противоречия в описанных системах, описаны итерации ТРИЗ-эволюции систем МП.

Таблица 1.4 – Спецификация итераций ТРИЗ-эволюции систем МП

№ п/п	**Переход**	**Итерация**	**Описание решения**
Перевод «слово-в-слово»			
1	Перевод «слово-в-слово» → Дословный МП	Con1.2: МР1↑, МР3↓ $\xrightarrow{IP:10}$ Sol1:МР3↑	Предварительно сформировать набор правил, в соответствии с которыми будет обрабатываться текст.
2	Перевод «слово-в-слово» → Трансферный МП	Con1.4: МР6↑, МР1↓ $\xrightarrow{IP:1}$ Sol2:МР6↓	Разбить сложное предложение на набор синтаксических структур.
3	Перевод «слово-в-слово» → Интерлингвистический МП	Con1.3: МР4↑, МР5↑ $\xrightarrow{IP:6,24}$ Sol3:МР5↓	Создать универсальный язык, переводить сначала на универсальный язык, а затем на целевой язык перевода. Для каждого языка прописать правила конвертации только

№ п/п	Переход	Итерация	Описание решения
			в одной языковой паре с промежуточным представлением.
4	Перевод «слово-в-слово» → Интерактивный МП	Con1.1: MP2↑, MP1.1↓ $\xrightarrow{\text{IP:15,16, 23}}$ Sol4:MP1↑	Переводить с погрешностью, предлагать пользователю в процессе перевода выбрать необходимый перевод слова, ориентируясь на контекст. Перестроить оставшуюся часть перевода в соответствии с внесенными пользователем изменениями.
		Con1.2: MP1↑, MP3↓ $\xrightarrow{\text{IP:16}}$ Sol5:MP1↑	Переводить как есть с погрешностями, связанными с грамматикой и стилистикой, готовый перевод пост-редактировать.
		Con1.4: MP6↑, MP1↓ $\xrightarrow{\text{IP:10, 24}}$ Sol6:MP6↓	Перед переводом либо в процессе перевода редактировать исходный текст таким образом, чтобы он стал более «понятным» для машины.
5	Перевод «слово-в-слово» → МП на примерах	Con1.1: MP2↑, MP1.1↓ $\xrightarrow{\text{IP:1,26,33}}$ Sol7:MP1.1↑	Разработать систему, которая будет делить текст на смысловые части, находить в базе и копировать перевод ранее переведенных частей, генерируя перевод только для нового фрагмента.
6	Перевод «слово-в-слово» → СМП по словам	Con1.3: MP4↑, MP5↑ $\xrightarrow{\text{IP:0}}$ Sol8: MP5.1↓	Вместо словарей в формате слово-слово обучать систему на параллельных корпусах в формате текст-эталонный перевод
7	Перевод «слово-в-слово» → СМП на основе синтаксиса	Con1.2: MP1↑, MP3↓ $\xrightarrow{\text{IP:10,17}}$ Sol9:MP3↑	Перед переводом производить полный синтаксический разбор предложения, структурируя части текста при помощи деревьев, затем переводить по словам.
8	Перевод «слово-в-слово» → СМП по фразам	Con1.1: MP2↑, MP1.1↓ & Con1.4: MP6↑, MP1↓ $\xrightarrow{\text{IP:1, 16, 17}}$ Sol10: MP1↑	Разбить предложение на фразы во всех возможных сочетаниях по n-слов в каждом и проанализировать перевод для каждого из них, выбрав в последствии только статистически наиболее верные сочетания, т.е. учитывая контекст.
Дословный МП			
9	Дословный МП → Интерактивный МП	Con2.1: MP1↑, MP7.1↑ $\xrightarrow{\text{IP:16}}$ Sol11:MP7.1↓	Переводить с погрешностью, предлагать пользователю скорректировать перевод с целью улучшения качества.
10	Дословный МП → МП на примерах	Con2.1: MP1↑, MP7.1↑ $\xrightarrow{\text{IP:2,26}}$ Sol12: {MP1.4↑, MP1.3↑, MP7.1↓}	Не описывать правила вообще, загрузить в систему большое количество примеров, из которых система

№ п/п	Переход	Итерация	Описание решения
			будет брать готовые части перевода.
11	Дословный МП → Трансферный МП	Con2.2: MP1↑, MP5↑ $\xrightarrow{IP:5,16}$ Sol13: MP5↓	Выделить ключевые синтаксические конструкции. Заложить в систему правила перевода каждого слова и подстановки в соответствии с синтаксическими конструкциями языка перевода.
12	Дословный МП → СМП по словам	Con2.2: MP1↑, MP5↑ $\xrightarrow{IP:2,24}$ Sol14: MP5↓	Создать промежуточный алгоритм обучения системы: на основе статистических методов система будет выбирать наиболее статистически вероятный перевода, при этом нет необходимости описывать правила.
		Con2.3: MP1↑, MP9↑ $\xrightarrow{IP:19}$ Sol15: MP9↓	Периодически подгружаем в систему новые параллельные корпуса текста, на которых система доучивается.
Трансферный МП			
13	Трансферный МП → МП на примерах	Con3.1: MP9↓, MP7.1↑ $\xrightarrow{IP:3,26}$ Sol16:MP7.1↓	Не описывать правила, загрузить в систему большое количество примеров, из которых система будет брать готовые части перевода.
14	Трансферный МП → СМП по словам	Con3.1: MP9↓, MP7.1↑ $\xrightarrow{IP:3,24}$ Sol17: MP7.1↓	Создаем промежуточный алгоритм обучения системы: на основе статистических методов система будет выбирать наиболее статистически вероятный вариант перевода, и не нужно описывать правила.
15	Трансферный МП → СМП по фразам	Con3.2: MP7.1↓, MP1.1↓ $\xrightarrow{IP:22}$ Sol18: MP1.1↑	Найти все возможные сочетания слов в тексте, при помощи статистических методов найти наиболее вероятный перевод.
16	Трансферный МП → Интерактивный МП	Con3.2: MP7.1↓, MP1.1↓ $\xrightarrow{IP:10,16}$ Sol19: MP1.1↑	Предварительно обрабатывать оригинал для того, чтобы он был более простым и «понятным» машине, переводить с погрешностью, привлекая пост-редактора для повышения качества перевода.
Интерлингвистический МП			
17	Интерлингвистический МП → СМП по словам	Con4.1: MP1↑, MP8↑ $\xrightarrow{IP:24,27}$ Sol20: MP8↓	Для редких языков и/или тематик из-за сложности создания и разметки универсальной модели и сбора обучающих данных использовать английский язык в качестве промежуточного.

№ п/п	Переход	Итерация	Описание решения
18	Интерлингвистический МП → Нейронный МП	Con4.1: MP1↑, MP8↑ $\xrightarrow{IP:17}$ Sol21: {MP1↑,MP8↓}	Для кодировки текста в универсальное представление использовать многослойную нейронную сеть.
МП на примерах			
19	МП на примерах → СМП по словам	Con5.1: MP1↑, MP7.2↑ $\xrightarrow{IP:10}$ Sol22:MP1.1↑	Перед копированием перевода выбрать из множества вариантов наиболее вероятный на основе статистических методов.
		Con5.2: MP10↑, MP1.4↓ $\xrightarrow{IP:10}$ Sol23:MP1.4↑	Записывать порядок слов в предложении и относительно друг друга, при переводе выбирать наиболее вероятный вариант на основе статистических методов
		Con5.4: MP1.5↑, MP1.2↓ $\xrightarrow{IP:11}$ Sol24:MP1.2↑	Заранее промаркировать места, где необходимо подставить служебное слово, при переводе заполнить маркеры.
20	МП на примерах → СМП по фразам	Con5.2: MP10↑, MP1.4↓ $\xrightarrow{IP:22}$ Sol25:{MP1.4↑, MP1.1↑}	Найти все возможные сочетания слов в тексте, при помощи статистических методов найти наиболее вероятный перевод.
		Con5.3: MP1.4↑, MP1.5↓ $\xrightarrow{IP:5,10}$ Sol26:MP1↑	При помощи статистических методов находить среди множества вариантов обучающего корпуса наиболее вероятный перевод целых фраз, а не отдельных слов.
21	МП на примерах → СМП на основе синтаксиса	{Con5.2: MP10↑, MP1.4↓, Con5.4: MP1.5↑, MP1.2↓} $\xrightarrow{IP:10,17}$ Sol27:{MP1.4↑, MP1.2↑}	Перед переводом производить полный синтаксический разбор предложения, структурируя части текста при помощи деревьев.
Интерактивный МП			
22	Интерактивный МП → Гибридный МП	Con6.1: MP1↑, MP11↑ $\xrightarrow{IP:0,3,5}$ Sol28:{MP1↑,MP11↓}	Использовать в качестве посредника между машинным переводом и человеком среду автоматизации переводов, объединив все цифровые инструменты, облегчающие перевод в одной среде. Использовать МП для тех фрагментов, для которых не нашлось совпадений по памяти перевода, при этом выбирать модель перевода исходя из особенности задачи.
Статистический МП по словам			
23	СМП по словам → СМП по фразам	Con7.1: MP1↑, MP5.2↑ $\xrightarrow{IP:5,16,17}$ Sol29:MP5.2↓	Объединить отдельные слова в предложении в фразы во всех возможных сочетаниях по n-слов в

№ п/п	Переход	Итерация	Описание решения
			каждом и проанализировать перевод для каждого из них, выбрав в последствии только статистически наиболее вероятные сочетания, т.е. учитывать контекст.
24	СМП по словам → Нейронный МП	Con7.2: MP4↑, MP1↓ $\xrightarrow{IP:6,17}$ Sol30: {MP1↑,MP4↑}	В качестве алгоритма вывода использовать нейронные сети, состоящие из энкодера и декодера. Энкодер сначала кодирует исходный текст в универсальное представление. Декодер распознает универсальное представление и переводит его на язык перевода.
Статистический МП по фразам			
25	СМП по фразам → Гибридный МП	{Con8.2: MP1.1↑, MP1.6↓, Con8.3: MP1.1↑, MP1.5↓} $\xrightarrow{IP:3,5}$ Sol31:MP1↑	Объединить в одной системе несколько типов систем, воспользовавшись преимуществами каждой из них, разделив задачи. Например: перевод отдельных слов по словарю и использование правил, а «выравнивание» текста доверить СМП или НМП, НМП переводить длинные предложения, СМП переводить короткие фразы.
26	СМП по фразам → Нейронный МП	{Con8.2: MP1.1↑, MP1.6↓, Con8.3: MP1.1↑, MP1.5↓} $\xrightarrow{IP:17,20}$ Sol32: MP1↑	Учитывать контекст всего предложения в переводе, а не только отдельных фраз при помощи многослойных нейронных сетей с долгосрочной памятью.
	СМП по фразам →	Con8.1: MP7↑, MP1.5↓	Решение отсутствует
Статистический МП на основе синтаксиса			
27	Syntax -based SMT → Нейронный МП	{Con9.1: MP1.2↑, MP12↑, Con9.2: MP4↑, MP1↓} $\xrightarrow{IP:2,\ 17}$ Sol33:MP1↑	Использовать многослойную нейронную сеть, не использовать методы синтаксического разбора и анализа.
Нейронный МП			
28	Нейронный МП → Гибридный МП	Con10.3: MP1.7↑, MP1.8↓ $\xrightarrow{IP:3,5}$ Sol34: {MP1.7↑, MP1.8↑}	Объединить в одной системе два типа систем, выбрав наиболее подходящую технологию отдельно для длинных предложений, отдельно для коротких фраз.
29	Нейронный МП → НМП «без учителя»	Con10.5: MP16↑, MP5↑ $\xrightarrow{IP:2}$ Sol35: MP5.2↓	Обучать сеть алгоритмом без учителя, исключив тем самым необходимость подготовки эталонной выборки данных.
30	Нейронный МП → Адаптированный НМП	Con10.6: MP16↑, MP1.9↓ $\xrightarrow{IP:5}$ Sol36:MP1.9↑	Для каждой из тематик готовить корпуса обучающих данных отдельно.

№ п/п	Переход	Итерация	Описание решения
	Нейронный МП →	Con10.1: MP1.6↑, MP13↑	Решение отсутствует
	Нейронный МП →	Con10.2: MP14↑, MP1.5↓	Решение отсутствует
	Нейронный МП →	Con10.4: MP1.6↑, MP15↑	Решение отсутствует
Адаптированный НМП			
31	Адаптированный НМП → НМП «без учителя»	Con11.1: MP1.9↑, MP5.2↑ $\xrightarrow{\text{IP:2, 17}}$ Sol37: MP5.2↓	Обучать нейронную сеть алгоритмом «без учителя» на моно-корпусах текста вместо параллельных корпусов.
НМП «без учителя»			
	НМП «без учителя» →	Con12.1: MP1↑, MP7↑	Решение отсутствует
Гибридный МП			
	Гибридный МП →	Con13.1: MP1↑, MP17↑	Решение отсутствует
	Гибридный МП →	Con13.2: MP16↑, MP8↑	Решение отсутствует
	Гибридный МП →	Con13.3: MP4↑, MP8↑	Решение отсутствует

Полная ТРИЗ-эволюционная карта представлена на рисунке 1.3.

Анализ ТРИЗ-эволюции систем МП [44, 45] показывает в первую очередь качественные скачки развития, которые позволили сформировать новые парадигмы: переход к трансферному МП привел к созданию парадигмы «перевод, основанный на правилах»; переход к СМП по словам – к созданию парадигмы «статистический МП»; переход к НМП – «к созданию парадигмы нейронный МП». По улучшаемым параметрам в ходе развития систем МП мы видим, что оно происходило по следующим ключевым направлениям: повышение качества перевода, причем с развитием МП и сам параметр качества трансформировался и детализировался в подпараметры; сокращение времени на сбор и подготовку обучающих данных; развитие способов обработки оригинала с целью упрощения его структуры для более точного понимания системой семантики текста; совершенствование технической реализации МП. Часть противоречий существующих систем не разрешены (8.1, 10.1, 10.2, 10.4, 12.1, 13.1, 13.2, 13.3), и, следовательно, задачи по разрешению этих противоречий являются перспективными направлениями исследования в области МП.

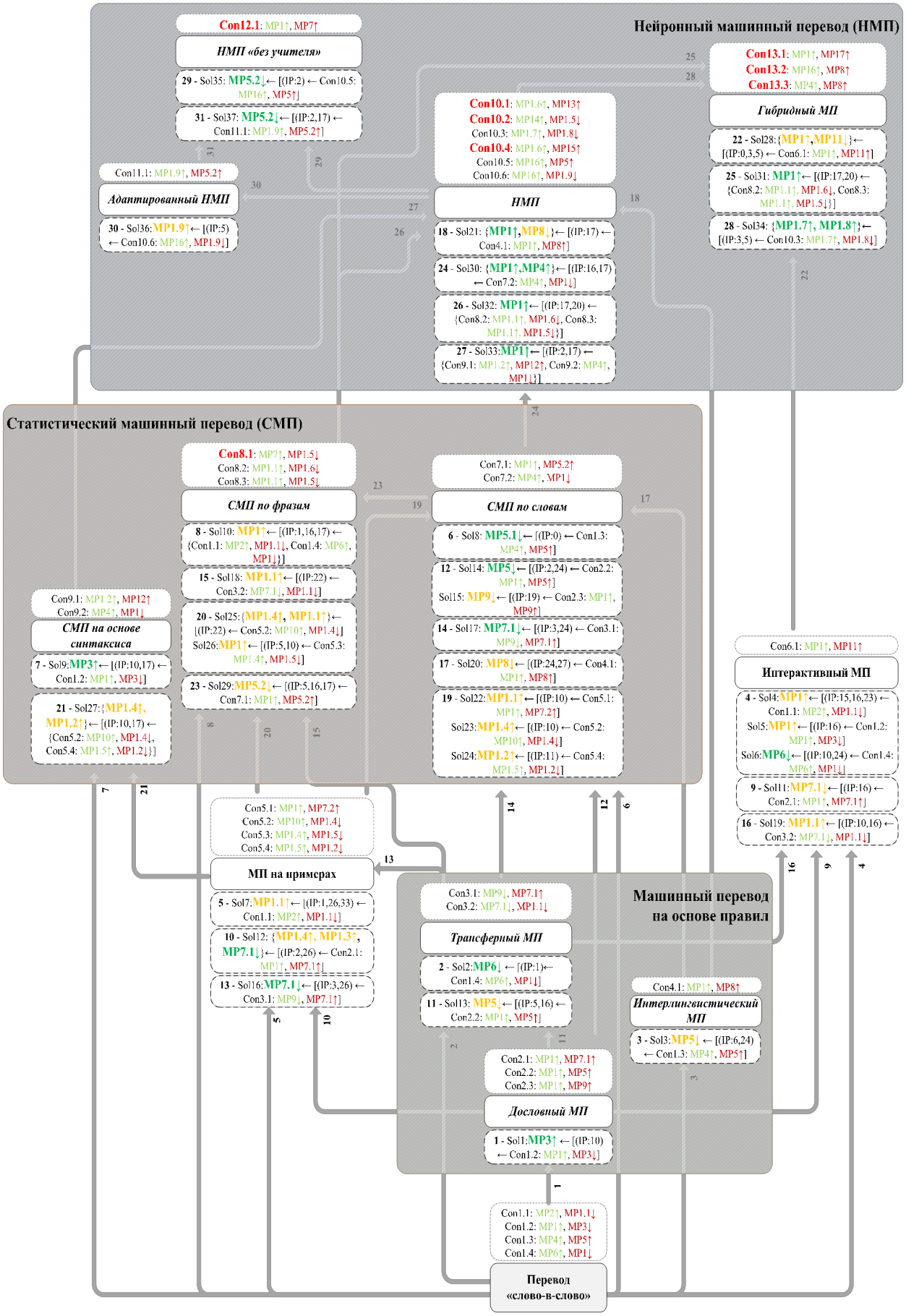

Рисунок 1.3 – ТРИЗ-эволюционная карта систем машинного перевода

По ТРИЗ-эволюционной карте и неразрешенным противоречиям мы видим, что одним из направлений совершенствования систем МП является **развитие методов интерактивного перевода** и нативного взаимодействия с пользователями МП. В системах МП не развиты методы обратной связи с пользователем, соответственно пользователь не может оценить риски некачественного перевода текста. из-за того, что система не может достаточно точно определять контекст текста целиком, а не отдельного предложения, оттенки смысла и т.д.

1.2 Интерактивный перевод и предредактирование

Несмотря на выдающийся прорыв МП в вопросах качества перевода, его применимость в прикладных задачах перевода все еще вызывает вопросы у пользователей и исследователей [46, 47], особенно при переводе узкоспециальных текстов [48]. В частности, C. Canfora и A. Ottmann в своем исследовании [49] поднимают вопрос рисков, связанных с высокими ожиданиями пользователей и индустрии переводов от внедрения технологий МП и ответственности за некачественный перевод в узкоспециальных предметных областях.

Очевидно, что для корректного использования систем МП необходимо прямое участие человека на всех этапах работы системы – от загрузки оригинала, сбора релевантных обучающих данных до оценки качества выполненного перевода [50]. Справедливо правило, что для достижения оптимального результата работы, пользователь системы МП должен знать хотя бы один язык из языковой пары перевода, и чем лучше знание языка перевода пользователя, тем точнее будет оценка качества МП и его пост-редактура. Другими словами, конечное качество перевода улучшается, если пользователь системы МП является носителем языка перевода.

Принимая во внимание этот принцип и рассматривая систему МП как средство автоматизированной поддержки работы переводчика, которое предоставляет переводчику модели и алгоритмы поддержки принятия решения относительно перевода, можно заключить, что, предоставляя пользователю средства обработки текста на языке, носителем которого он является, на любом из этапов перевода, можно

повысить качество перевода. Ключевая идея данного подхода [51] в том, что на разных стадиях перевода привлекается участие человека. Такое участие может быть выражено в разных формах:

- *с постредактированием*: человек редактирует результат МП;
- *с предредактированием*: человек редактирует входной текст, приспосабливая его для более легкого понимания машиной;
- *частично автоматизированный перевод*: человек и машина взаимодействуют в процессе перевода. Например, использование человеком электронных словарей при переводе; участие человека в процессе машинного перевода для разрешения трудностей.

Работы по частично-автоматизированному переводу [52, 53, 54] обычно рассматривают варианты совершенствования автоматизированного инструментария переводчиков с целью повышения скорости и качества их работы.

Множество работ и исследований посвящено постредактированию машинного перевода. В литературе описываются стратегии постредактирования, вопросы подготовки постредакторов, типичные ошибки машинного перевода, автоматизации постредактирования т.д. [55, 56, 57, 58]. Предредактированию посвящено меньше работ, однако существующие практические статьи указывают на эффективность такого подхода к повышению качества машинного перевода [59, 60].

Seretan, Bouillon и Gerlach в своем исследовании [61] показали, что использование даже простых полуавтоматических правил предредактирования текста повышает качество МП. Гипотеза была проверена на сообщениях пользователей технического форума и медицинских текстах и показала одинаковую эффективность. Авторы указывают, что реализация методов автоматического предредактирования возможна только для языков, для которых уже созданы поверхностные синтаксические анализаторы.

Ряд исследований сравнивают подходы постредактирования и предредактирования. Так, исследование C. Shei [62] в языковой паре китайский-английский показало преимущества и позитивное влияние предредактирования в сравнении с

постредактированием на качество МП для пользователей с низким уровнем владения языком перевода.

Y. Liang и W. Han в задаче перевода медицинских текстов [63] сравнили оба подхода по критериям снижения количества ошибок НМП и сохранения экономической эффективности. Результаты показали, что предредактирование повышает качество перевода и позволяет повысить экономическую эффективность по сравнению с пост-редактированием.

Gerlach, Porro, Bouillon и Lehmann [64] показали, что в прикладных задачах оба подхода могут использоваться совместно, при этом предредактирование позволяет сократить время пост-редактуры МП и в 65% случаев повышает качество результата МП.

Исследование, проведенное коллективом российских ученых в задаче перевода новостей с английского языка на русский [65], также подтвердило, что предредактирование исходного текста сокращает время пост-редактуры МП на 30-40%.

Несмотря на подтвержденную исследования эффективность предредактирования как способа повышения качества перевода, следует учитывать и тот факт, что при автоматизации эффект от предварительного редактирования может быть непредсказуемым, включая отсутствие положительного эффекта и даже преобладание отрицательного [66].

A. Arenas в широкомасштабном анализе обоих подходов [67] показала, что способ предредактирования и методология его применения должны адаптироваться под особенности языковой пары перевода, специфику предметной области, а также используемую технологию МП. Отмечается, что необходим анализ исходного текста с целью определить «уязвимости» с точки зрения применяемой технологии МП, на основе которого становится возможной выбор стратегии предварительного редактирования.

1.3 Методы предредактирования для повышения качества машинного перевода

1.3.1 Правила для авторов по написанию исходных текстов

Один из способов повышения качества МП заключается в том, чтобы изначально написать текст «правильно», минимизировав вероятность полисемии и типичных ошибок [68, 69]. Каждый язык имеет грамматические правила. Не существует таких правил, которые давали бы подходящие для всех языков результаты. Однако существуют правила, которые снижают уровень двусмысленности в большинстве текстов на многих языках, например, «писать короткими и грамматически простыми предложениями», «использовать существительные вместо местоимений», «использовать определяющие слова», «использовать активный залог вместо пассивного».

Это позволяет создавать тексты, которые легче читать, понятнее и легче запоминать, а также с лучшим словарным запасом и стилем. В связи с этим многие компании, использующие МП, а так же разработчики систем МП разрабатывают и выпускают различные рекомендации для пользователей МП по написанию исходных текстов [70, 71]. Существенным недостатком данного подхода является то, что автор исходного текста и пользователь системы МП – не всегда один и тот же человек, поэтому зачастую авторы игнорируют эти рекомендации, так как не заинтересованы в получении перевода написанного ими текста [72].

1.3.2 Концепция контролируемого языка

Контролируемый язык – это подвид естественного языка, полученный ограничением в использовании грамматики, терминологии и речевых оборотов посредством регламентирующих правил с тем, чтобы снизить или искоренить его многозначность и сложность [73].

Традиционно контролируемые языки подразделяются на две группы: в одной все усилия направлены на повышение удобочитаемости для человека (например, для тех, кому язык текста не родной) [74]; в другой эти меры направлены на создание языка, надежного в плане автоматического семантического анализа, в частности, для повышения эффективности систем МП [75].

Контролируемый язык особенно полезен в задачах машинного перевода, где требуется высокая точность и аккуратность, например, перевода юридических документов, медицинских текстов или технических спецификаций [76], так как позволяет получить более единообразный и стандартизированный исходный текст, что обеспечивает как более высокую частоту совпадений в памяти переводов, так и более высокое качество генерируемого МП. Однако, несмотря на преимущества концепции контролируемого языка, она сложна и трудоемка в реализации и внедрении.

1.3.3 Предредактирование, основанное на правилах

В отличии от правил для авторов и концепции контролируемого языка, предредактирование предполагает, что редактор подготавливает и оптимизирует исходный текст под задачу МП. Выделяют следующие принципы предредактирования:

- Удаление неоднозначных слов и выражений, которые могут иметь несколько значений на языке перевода.
- Трансформация сложных грамматических конструкций в более простые, к которым проще найти соответствие в памяти переводов и которые проще перевести.
- Замена терминологии на стандартизированную, расшифровка аббревиатур, что помогает обеспечить согласованность перевода и избежать ошибок.

Так, Hiraoka и Yamada [77] в своей работе предприняли попытку сформулировать основные правила предредактирования текста при переводе с японского языка на английский. При этом, они разделяют предредактирование на два вида: одноязычное и двуязычное. При одноязычном предредактировании редактор знает только язык оригинала и, следуя рекомендация или «правилам» вносит правки с исходный текст. При двуязычном предредактировании редактор оценивает оригинальный текст, выполненный машинный перевод и корректирует оригинальный текст с целью добиться желаемого перевода. Их стратегия показала статистически значимые результаты, в том числе и для китайского и корейского языков. Подобные исследования проводились для перевода с японского на английский и восточные языки [78], с индонезийского на английский [79], с английского на испанский [80]. Все указанные выше исследования показали, что предредактирование и переписывание исходного текста, опираясь на правила, повышает качество МП. В качестве

недостатков такого подхода можно выделить трудоемкость, зависимость результата от профессионализма редактора.

1.3.4 Решение задачи перефразирования как способ предредактирования

Разработки в области обработки естественного языка и возросшая доступность текстовых корпусов сделали возможным развитие методов и алгоритмов автоматической адаптации (упрощения) текста, которые могут быть использованы и в задачах МП [81]. Одним из перспективных направлений в этой области является автоматическое перефразирование текста [82] при котором создаются новые версии текста с сохранением семантики исходного текста. Путем изменения фраз, предложений или структуры текста текст адаптируется текст под требования реципиента [83].

Существует несколько подходов к решению задачи автоматического перефразирования текстов. Один из них основывается на использовании правил и шаблонов для замены слов или фраз на эквивалентные им варианты. Несмотря на активное изучение проблемы, оптимальные решения пока найдены далеко не для всех задач в этой области, а существующие методы, применимы к сравнительно небольшому числу языков [84, 85, 86, 87]. Для русского языка разработан ряд правил и алгоритмов по упрощению синтаксической сложности [88] и многозначности отдельных частей речи [89].

Другой подход – это использование методов глубокого обучения, таких как рекуррентные нейронные сети, которые позволяют модели генерировать новые версии текста на основе обучающих данных [90]. Это сложная задача, которая требует понимания контекста и семантики исходного текста, а также умения создавать новые, грамматически правильные и естественно звучащие предложения [91]. Анализу русскоязычных текстов посвящены работы и автоматизации оценки восприятия текста посвящены ряд работ [92, 93], однако они не предлагают алгоритмов генерации текста на основе выявленных семантик. Кроме того, для обучения моделей перефразирования требуются корпуса параллельных текстов большого объема, которые сложно собрать, особенно в узкоспециальных и высоко конкурентных предметных областях.

Выводы по первой главе

Анализ существующих исследований по теме автоматизированной предобработки исходных текстов для МП, что является перспективным направлением исследований в соответствии с определенными ключевыми направлениями развития систем МП, описанными выше, показал, что:

- не проводилось исследований по анализу ошибок нейронного МП с русского языка, связанных со спецификой той или иной предметной области или с особенностями языка;
- несмотря на то, что для русского языка разработаны синтаксические анализаторы и возможна разработка правил автоматического предредактирования для повышения качества МП, существующие исследования носят точечный несистемный характер;
- в литературе незначительно описаны правила и алгоритмы автоматического предредактирования текстов в контексте задачи повышения качества МП, существующие прикладные исследования ориентированы в большей степени на адаптацию текстов для реципиентов различной квалификации;
- существующие подходы к автоматизации предредактирования имеют ограничения, связанные со сбором и подготовкой обучающих данных.

Обобщая вышесказанное можно заключить, что являются актуальными задачи определения зависимости качества перевода от параметров исходного текста и стратегий его обработки с целью максимизации качества машинного перевода, а также разработки методов и алгоритмов оптимизационного предредактирования текстов на русском языке согласно особенностям языков оригинала и перевода, специфичных для определенной предметной области, и технологии МП.

ГЛАВА 2 МАТЕМАТИЧЕСКАЯ МОДЕЛЬ ОПТИМИЗАЦИОННОГО ПРЕДРЕДАКТИРОВАНИЯ

2.1 Обобщенная математическая модель процесса перевода

Решение задачи разработки моделей, методов и комплекса программ оптимизационного предредактирования текстов для дальнейшего перевода, лежит в области прикладной компьютерной лингвистики.

Задача компьютерной лингвистики (КЛ) – разработка методов и средств построения лингвистических процессоров для прикладных задач по автоматической обработке текстов на естественном языке (ЕЯ) [94]. Разработка лингвистического процессора для некоторой прикладной задачи предполагает формальное описание лингвистических свойств обрабатываемого текста, которое может рассматриваться как модель текста (или модель языка).

Машинный перевод – самое раннее приложение КЛ, вместе с которым возникла и развивалась сама эта область. Первые программы перевода были построены на простейшей стратегии пословного перевода, однако довольно быстро стало ясно, что машинный перевод требует гораздо более полной лингвистической модели.

Компьютерная лингвистика тесно связана с областью искусственного интеллекта (ИИ) [95], в рамках которой разрабатываются программные модели отдельных интеллектуальных функций. Несмотря на очевидное пересечение исследований в области компьютерной лингвистики и ИИ (поскольку владение языком относится к интеллектуальным функциям), ИИ не включает в себя всю КЛ, поскольку она имеет свой теоретический базис и методологию. Общим для указанных наук является компьютерное моделирование как основной способ и итоговая цель исследований, эвристический характер многих применяемых методов.

Машинное обучение в КЛ применяется для обработки коллекций текстовых документов, с использованием признаковой модели текста, при которой признаки определены для каждого документа по отдельности. Признаками могут выступать различные информационные характеристики текста: как лингвистические, так статистические и структурные: например, частота определенных слов (или их

категорий) в документе, частота использования спецзнаков, соотношение частей речи слов, наличие определенных синтаксических конструкций или разделов текста, дата создания и прочие [94].

Для постановки и формализации задачи настоящего исследования, проанализируем процесса перевода текста с одного естественного языка на другой с декомпозицией на этапы и построим обобщенную математическую модель [96]. В качестве основной методологии описания математической модели используем теорию множеств.

2.1.1 Термины и определения

Определения элементов модели сформулированы на основании анализа предметной области и фундаментальных исследований в области лингвистики [97].

Язык – сложная знаковая система, естественно или искусственно созданная и соотносящая понятийное содержание и типовое звучание (написание).

Множество существующих языков можно представить следующим образом:

$$яз \in ЯЗ: ЯЗ=\{яз_0, яз_1, \dots яз_{nЯЗ}\}. \quad (3)$$

Понятийное содержание определяется семантическими единицами.

Семантическая единица – это абстрактная сущность, не привязанная к конкретному языку или иному средству выражения (например, языку математической логики), которая обеспечивает идентификацию определенных объектов, явлений, свойств домена приложения.

Пример: Семантическая единица – дождь. Способы выражения: дождь; атмосферные осадки в виде водяных капель; природное явление, когда с неба падают капли воды; rain (англ.); regen (нем.); и т.д.

Множество семантических единиц можно представить следующим образом:

$$ce \in CE: CE=\{ce_0, ce_1, \dots ce_{nCE}\}. \quad (4)$$

Пусть множество U_{ce} – универсальное множество всех возможных описаний на всех возможных языках всех возможных семантических единиц. *Ice* – множество индексов-количество семантических единиц:

$$ice \in Ice: Ice = \{0, 1, \dots, nCE\}. \quad (5)$$

При этом множества *CE* и *Ice* являются биективными, т.е. $Ice \leftrightarrow CE$.

Каждому элементу $ice \in Ice$ однозначно сопоставлено подмножество описаний семантической единицы $O_{ice} \subseteq U_{ce}$. Тогда $CoCE = (O_{ice})_{ice \in Ice}$ – счетное семейство множеств описаний семантических единиц.

$$CoCE = (O_{ice})_{ice \in Ice} = \bigcup_{ice \in Ice} O_{ice} = \{x: \exists ice\ x \in O_{ice}\} \quad (6)$$

Домен приложения или **предметная область** – абстрактное понятие, которое определяет множество объектов в пределах общего контекста. Множество доменов (предметных областей) приложения:

$$дп \in ДП: ДП = \{дп_0, дп_1, \ldots дп_{nДП}\}. \quad (7)$$

Один из способов выражения семантических единиц и коммуникации в рамках некоторого домена приложения – текст.

Текст – это письменное сообщение, объективированное в виде письменного документа, состоящее из ряда высказываний, объединённых разными типами лексической, грамматической и логической связи, имеющее определённый моральный характер, прагматическую установку и соответственно литературно обработанное. Текст любого размера – это относительно автономное (законченное) высказывание. Множество текстов:

$$txt \in TXT: TXT = \{txt_0, txt_1, \ldots txt_{nTXT}\}. \quad (8)$$

Уровни фрагментации текста: Символ → Морфема (корень, суффикс и т.д.) → Лемма → Словоформа → Лексема → Словосочетание → N-грамма → Предложение → Абзац → Раздел → Документ (книга) → Корпус.

Текст как объект анализа и исследования обладает некоторым набором свойств или признаков. Множество признаков или свойств текста:

$$св \in СВ: СВ = \{св_0, св_1, \ldots св_{nСВ}\}. \quad (9)$$

В задачах обработки естественного языка свойства текста условно можно разделить на группы признаков: *общие* (количество символов/слов/строк и т.д., стиль, язык, домен приложения и пр.), $ОП \subset СВ$; *лексические* (процент покрытия текста лексическими минимумами, частотными списками и др.), $ЛП \subset СВ$; *морфологические* (количество различных частей речи и грамматических форм), $МП \subset СВ$; *синтаксические* (глубина глагольных и именных групп, связи между глаголами в

предложениях), *СП ⊂ СВ*; *признаки, основанные на базовых подсчетах* (средняя длина слов и предложений и пр.), *БП ⊂ СВ*.

$$СВ = ОП \cup ЛП \cup МП \cup СП \cup БП \quad (10)$$

Совокупность свойств и признаков определяет главные параметры текста, к которым относится целостность, связность, трудность, удобочитаемость, сложность и другие. Множество главных параметров текста:

$$гп \in ГП: ГП = \{гп_0, гп_1, \ldots гп_{nГП}\}. \quad (11)$$

Основные параметры текста – удобочитаемость и сложность.

Удобочитаемость позволяет оценить, соответствует ли текст читательской способности реципиента, а на уровне отдельных его единиц (например, предложений) выявить элементы, требующие упрощения (специальные термины, аббревиатуры, сокращения, отсутствие контекста (таблицы, счета, картинки и т. д.), слишком короткие или слишком длинные предложения, ошибки и пр.). Для оценки удобочитаемости разработаны различные метрики: индекс Флеша, школьный тест Флеша-Кинкейда, FOG, SMOG, Индекс Коулман-Лиау, Автоматический индекс удобочитаемости [98].

Сложность текста определяется морфологическими, лексическими, синтаксическими признаками текста (объективные оценки), а также его информационной (когнитивной) трудностью (субъективная оценка). Когнитивная трудность текста определяется на уровне реципиента его способностью идентифицировать семантические единицы в тексте [99].

Перевод – деятельность по интерпретации понятийного содержания текста на одном языке и созданию нового текста на другом языке эквивалентного понятийного содержания.

Перевод выполняется переводчиком из множества переводчиков, компетенция каждого из которых определяется набором знаний, умений и навыков в области владения языками. Множество переводчиков:

$$пер \in ПЕР: ПЕР = \{пер_0, пер_1, \ldots пер_{nПЕР}\}. \quad (12)$$

Знания языков или **языковая компетенция** – это владение грамматическими и словарными аспектами языка. Знания о языках можно охарактеризовать

совокупностью множеств лексем, грамматических правил, морфологических правил, правил синтаксиса и пунктуации, словарей частотности словоупотребления.

Лексема – абстрактная двусторонняя единица словарного состава языка в совокупности всех ее конкретных грамматических форм и выражающих их флексий, а также всех возможных значений во всех его употреблениях и реализациях.

Множество лексем определяется множеством лемм:

$$лм \in ЛМ: ЛМ = \{лм_0, лм_1, \ldots лм_{nЛМ}\}, \quad (13)$$

где $лм_{iЛМ}$ – начальная словарная форма слова или лемма. В русском языке для существительных и прилагательных это форма именительного падежа единственного числа, для глаголов и глагольных форм – форма инфинитива.

Пусть множество $U_{сфл}$ – универсальное множество всех возможных словоформ всех возможных лемм на всех возможных языках, а $Iлм$ – множество индексов-количество лемм:

$$iлм \in Iлм: Iлм = \{0, 1, \ldots, nЛМ\}, Iлм \leftrightarrow ЛМ \quad (14)$$

Каждому элементу $iлм \in Iлм$ однозначно сопоставлено подмножество лексем $ЛКС_{iлм} \subseteq U_{сфл}$. Тогда $СмЛКС = (ЛКС_{iлм})_{iлм \in Iлм}$ – счетное семейство множеств лексем, т.е.

$$СмЛКС = (ЛКС_{iлм})_{iлм \in Iлм} = \bigcup_{iлм \in Iлм} ЛКС_{iлм} = \{x: \exists iлм\ x \in ЛКС_{iлм}\}. \quad (15)$$

Количество словоупотреблений одной леммы во всех вариантах ее лексем в некотором тексте (или в корпусе текстов, или в речевом фрагменте) называется ее частотностью, измеряется в лексической статистике и описывается в **частотных словарях**. Пусть множество $U_{част}$ – универсальное множество всех частотностей всех возможных лемм и сочетаний слов на всех возможных языках для всех доменов приложения. Каждому элементу $iлм \in Iлм$ однозначно сопоставлено подмножество частотностей $ЧАСТ_{iлм} \subseteq U_{част}$. Тогда $СмЧАСТ = (ЧАСТ_{iлм})_{iлм \in Iлм}$ – счетное семейство множеств частотностей употребления лемм.

$$СмЧАСТ = (ЧАСТ_{iлм})_{iлм \in Iлм} = \bigcup_{iлм \in Iлм} ЧАСТ_{iлм} = \{x: \exists iлм\ x \in ЧАСТ_{iлм}\} \quad (16)$$

Интерпретация и перевод слов на другие языки содержатся в словарях. **Словарь** – это собрание слов, устойчивых выражений с пояснениями, толкованиями и/или с переводом на другой язык, словарь состоит из словарных статей.

Словарная статья – это кортеж сочетаний слов (словоформ отдельных лемм) длинной k и n ($k, n \leq 4$) на разных языках, таких что описывают максимально близко друг другу одну семантическую единицу, то есть

$$СС = \{(соч.\ слов\ 1;\ соч.\ слов\ 2): се_{iCE} \mid соч.\ слов\ 1 \rightarrow се_{jCE} \mid соч.\ слов\ 2\}, \quad (17)$$

$$соч.\ слов\ 1 = (слЯ31_{0сф}, слЯ31_{iсф}, \ldots, слЯ31_{mсф}), \quad (18)$$

$$соч.\ слов\ 2 = (слЯ320сф, слЯ32jсф, \ldots, слЯ32nсф), \quad (19)$$

$$слЯ31_{iсф}, слЯ32_{jсф} \in U_{сфл}, \quad (20)$$

$$се_{iCE}, се_{jCE} \in CE. \quad (21)$$

Пусть множество $U_{сс}$ – универсальное множество всех возможных словарных статей. Каждому элементу множества индексов количества семантических единиц $ice \in Ice$ однозначно сопоставлено множество словарных статей $СЛ_{ice} \subseteq U_{сс}$, то есть $СмСЛ = (СЛ_{ice})_{ice \in Ice}$ – счетное семейство множеств словарей.

$$СмСЛ = (СЛ_{ice})_{ice \in Ice} = \bigcup_{ice \in Ice} СЛ_{ice} = \{x: \exists ice\ x \in СЛ_{ice}\} \quad (22)$$

Правила употребления слов можно охарактеризовать набором следующих множеств:

- множество грамматических правил:

$$гп \in ГП : ГП = \{гп_0, гп_1, \ldots гп_{nГП}\}; \quad (23)$$

- множество морфологических правил:

$$мор \in МОР: МОР = \{мор_0, мор_1, \ldots, мор_{nМОР}\}; \quad (24)$$

- множество правил синтаксиса и пунктуации:

$$син \in СИН : син = \{син_0, син_1, \ldots син_{nСИН}\}. \quad (25)$$

Умения переводчика – это способности переводчика, которые позволяют ему интерпретировать смысловую единицу, выраженную в тексте на языке оригинала, и создать эквивалентный по смыслу текст на языке перевода в соответствии с предъявляемыми к переводу требованиями. Множество умений:

$$ум \in УМ: УМ = \{ум_0, ум_1, \ldots ум_{nУМ}\}, \quad (26)$$

где $ум_{iУМ}$ – функция обработки, преобразования или генерации текста таким образом, чтобы созданный текст максимально точно передавал смысл исходного текста, то есть

$$ум_{iУМ}\ (txt_{iTXT}) = txt'_{iTXT} : CM|txt'_{iTXT} \rightarrow CM|txt_{iTXT}\ , \quad (27)$$

где СМ – **смысл** или упорядоченный набор семантических единиц, описываемый текстом:

$$CM = \{(ce_0, ce_1, \ldots, ce_{псм}): ce_{iсм} \in CE\}. \quad (28)$$

Множество всех смыслов:

$$CM \in мCM: мCM = \{CM_0, CM_1, \ldots CM_{nCM}\}. \quad (29)$$

Умения связаны со способностью выполнять операции над текстом согласно требованиям качества. Множество требований к качеству перевода можно описать следующим образом:

$$mp \in TP: TP=\{mp0, mp1, \ldots mpnTP\}, \quad (30)$$

где mp_{iTP} – это некоторое лингвистическое требование, которому должен соответствовать конечный результат перевода, может быть записано в виде условия либо утверждения.

Требования определяют: нормы языка перевода $НЯ \subset TP;$ культурные традиции реципиента $КТР \subset TP;$ отраслевые стандарты согласно домену приложения $ОС|дп_{iДП} \subset TP;$ заказчик и/или реципиент перевода $ЗР \subset TP.$

Множество *ТР* включает и все прочие требования, не вошедшие в другие подмножества требований. Подмножества требований могут пересекаться между собой. Требования к переводу от заказчика и/или реципиента – это множество сочетаний слов, используемых для выражения семантических единиц в текстах заказчика. При наличии требований от заказчика, частотность сочетаний слов, описанных в таких требованиях, при переводе текстов заказчика принимается как максимальная.

Навыки переводчика – это опыт использования языка в решении задач перевода в различных предметных областях, полученный в результате обучения либо практической деятельности. Навыки характеризуются накоплением опыта через следующие понятия.

Множество **устойчивых n-грамм**, т.е. наборов слов неограниченной длины, которые описывают некоторые смыслы (наборы семантических единиц). Взаимосвязи внутри наборов семантических единиц и их составы, а также n-граммы описывающие составленные наборы, которые обобщенно можно назвать текстами $txt \in TXT$, для различных доменов приложения определяются / дополняются / корректируются переводчиком в процессе обучения и/или практической деятельности, то есть

$$усоч \in УСОЧ: УСОЧ=\{усоч_0, усоч_1, \dots усоч_{nУСОЧ}\}, \quad (31)$$

где

$$усоч_{nУСОЧ}= \{(txt, СМ): txt \in TXT, СМ \in мСМ\}. \quad (32)$$

Семейство множеств **частотности употребления устойчивых n-грамм** для выражения смыслов вообще и/или в рамках определенного домена. Пусть *Iусоч* – множество индексов-количество устойчивых n-грамм:

$$iусоч \in Iусоч: Iусоч = \{0, 1, \dots, nУСОЧ\}, Iусоч \leftrightarrow УСОЧ. \quad (33)$$

Каждому элементу $iусоч \in Iусоч$ однозначно сопоставлено подмножество частотностей $УСЧАСТ_{iусоч} \subset U_{част}$. Тогда $СмУСЧАСТ = (УСЧАСТ_{iусоч})_{iусоч \in Iусоч}$ – счетное семейство множеств частотностей употребления устойчивых n-грамм.

$$СмУСЧАСТ =(УСЧАСТ_{iусоч})_{iусоч \in Iусоч} =$$

$$= \bigcup_{iусоч \in Iусоч} ЧАСТ_{iусоч}= \{x: \exists iусоч\ x \in УСЧАСТ_{iусоч}\} \quad (34)$$

Специализация переводчика $пер_{iПЕР}$ – это его накопленный опыт выражения семантических единиц в рамках некоторой предметной области $дп_{iДП}$, включая обучение. Математически определим специализацию переводчика в рамках решения прикладной задачи как вектор, координатами которого являются условные подмножества:

$$\overrightarrow{Спер_{iПЕР},дп_{iДП}} = \{(ОТ|дп_{iДП})|пер_{iПЕР}, (УСОЧ|дп_{iДП})|пер_{iПЕР},$$
$$(СмУСЧАСТ|\ дп_{iДП})|пер_{iПЕР}\} \quad (35)$$

Компетенция некоторого переводчика $пер_{iПЕР}$ оценивается всегда в рамках прикладной задачи, исходными данными которой являются язык оригинала $яз_{вх} \in ЯЗ$ и язык перевода $яз_{вых} \in ЯЗ$. Математически определим компетенцию переводчика в рамках решения прикладной задачи как вектор, координатами которого являются условные подмножества:

$$\overrightarrow{Кпер_{iПЕР}} = \{(СоСЕ\,|\,яз_{вх} \cup СоСЕ\,|\,яз_{вых})|пер_{iПЕР}, \quad (36)$$

$$(СмЛКС|\,яз_{вх} \cup СмЛКС|\,яз_{вых})|пер_{iПЕР},$$

$$(СмЧАСТ|яз_{вх} \cup СмЧАСТ|яз_{вых})|пер_{iПЕР},$$

$$(СмСЛ|яз_{вх} \cup СмСЛ|яз_{вых})|пер_{iПЕР},$$

$$(ГП|яз_{вх} \cup ГП|яз_{вых})|пер_{iПЕР},$$

$$(МОР|яз_{вх} \cup МОР|яз_{вых})|пер_{iПЕР},$$

$$(СИН|\,яз_{вх} \cup СИН|яз_{вых})|пер_{iПЕР},$$

$$(УМ|яз_{вх} \cup УМ|яз_{вых})|пер_{iПЕР},$$

$$(УСОЧ|яз_{вх} \cup УСОЧ|яз_{вых})|пер_{iПЕР},$$

$$(СмУСЧАСТ|яз_{вх} \cup СмУСЧАСТ\,|яз_{вых})|пер_{iПЕР}\}$$

Рассмотрим применение описанных элементов модели в процессе перевода.

2.1.2 Оценка задачи перевода

Переводчик $пер_{iПЕР}$ получает текст txt_{iTXT} на языке $яз_{вх}$ для перевода на язык $яз_{вых}$ и требования к переводу от заказчика $ЗР|txt_{iTXT}$ (если есть), при этом $ЗР|txt_{iTXT} \subset ТР|txt_{iTXT}$.

Перед выполнением перевода переводчик оценивает **сложность задачи перевода**. При оценке сложности задачи перевода переводчик обращает внимание на неизвестные ему слова и сочетания слов на языке $яз_{вх}$, для которых он не может идентифицировать значение смысловой единицы, либо смысловые единицы, для которых он не может найти аналог на языке перевода $яз_{вых}$ среди известных ему слов и сочетаний слов. Множества свойств и параметров исходного текста $СВ|txt_{iTXT}$ и $ГП|txt_{iTXT,}$ и то, обладает ли переводчик достаточной компетенцией $\overrightarrow{Кпер_{iПЕР}}$ относительно языков $яз_{вх}$ и $яз_{вых}$ и специализацией $\overrightarrow{Спер_{iПЕР,}дп_{iДП}}$, т.е. навыками описания семантических единиц на языке перевода в рамках заданной предметной области исходного текста, определяет вероятность создания переводчиком переведенного текста на таком уровне качества, который определяется требованиями $ТР|txt_{iTXT}$.

Алгоритм оценки сложности задачи перевода следующий:

Шаг 1. Исходя из домена приложения текста $дп_{iДП}$, формируется множество оценок текста $ОЦ = СВ \cup ГП$.

Шаг 2. Для каждого значения $св_{iСВ}, гп_{iГП} \in ОЦ$, на основе требований к переводу $ТР|txt_{iТХТ}$, компетенций переводчика относительно языковой пары $\overrightarrow{Кпер_{iПЕР}}$ и специализации переводчика относительно домена приложения текста $\overrightarrow{Спер_{iПЕР,}дп_{iДП}}$ формируется значение значимости $w_{оцk}$, множество нормированных значений $w_{оцk}$ значимости формируют матрицу значимости оценок сложности $W_{оц}$ размерностью $1\times k$, где k – общее число оценок, которые выступают коэффициентами уравнения поиска теоретического значения качества перевода.

Шаг 3. Для каждого i-го фрагмента текста при $i=\overline{1,N}$ формируется матрица оценок фрагмента исходного текста $С_{оцi}$ размерностью $1\times k$, где k – общее число оценок.

Шаг 4. На основании оценок $С_{оцi}$ значимости $W_{оц}$ формируется уравнение поиска теоретического результирующего фактора, т.е. качества перевода $\widehat{ОК}$:

$$\widehat{ОК}_i(С_{оцi}, W_{оц}) = w_0 + w_{1оц}С_{оцi_1} + w_{2оц}С_{оцi_2} + \cdots + w_{kоц}С_{оцik} \quad (37)$$

Шаг 5. Для каждого *i-го* фрагмента текста рассчитаем вероятность получения переведенного текста на таком уровне качества, который определяется требованиями $ТР|txt_{iТХТ}$, применив к уравнению (37) логит-преобразование [100]:

$$p_i = \frac{1}{1+e^{-\widehat{ОК}_i}} \quad (38)$$

Шаг 6. Сложность задачи перевода *i-го* фрагмента текста:

$$СлЗП_i = \frac{1}{p_i}, \quad (39)$$

где p_i – это вероятность создания переводчиком перевода требуемого качества, рассчитанная по формуле (38).

Шаг 7. Результирующая сложность задачи перевода текста – это наибольшее значение сложности задачи перевода $СлЗП_i$ среди N фрагментов исходного текста:

$$СлЗП_{txtiТХТ} = max\, СлЗП_i \quad (40)$$

В зависимости от значения $СлЗП_{txt_{iТХТ}}$ переводчик выбирает стратегию дальнейшей обработки текста. Возможные варианты для ручного и машинного перевода приведены в таблице 2.1.

Таблица 2.1 – Действия переводчика в зависимости от сложности задачи перевода

Сложность задачи перевода	**Действие**	
	Ручной перевод	*Машинный перевод*
Низкая	Перевод	Перевод
Средняя	Предварительная редактура текста / согласование требований к переводу с заказчиком	Автоматическая предредактирование текста
Высокая	Отказ от перевода / отправка на доработку текста заказчику	Полуавтоматическая предредактирование текста с привлечением пользователя

Диапазоны значений *СлЗП*, соответствующие низкому, среднему и высокому уровням определяются на основе предварительного анализа требований к переводу и с учетом способа перевода (ручной/машинный).

Формально функцию оценивания сложности переводческой задачи $F_{оц}$ можно представить следующим образом:

$$F_{оц} : txt_{iTXT} \rightarrow СлЗП|(ТР|txt_{iTXT,}\ СВ,\ ГП,\ \overrightarrow{Кпер_{iПЕР}},\ \overrightarrow{Спер_{iПЕР,}дп_{iДП}}) \quad (41)$$

Схематично, функция оценки сложности переводческой задачи представлена на рисунке 2.1.

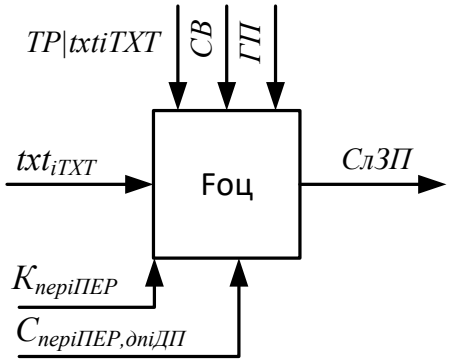

Рисунок 2.1 – Оценка сложности задачи перевода

2.1.3 Предредактирование исходного текста

Предредактирование – это процесс автоматического/полуавтоматического изменения свойств/параметров исходного текста с целью оптимизации сложности переводческой задачи с учетом влияния такого изменения на качество перевода.

Предредактирование выполняется при средней и высокой сложности переводческой задачи с использованием разнообразных алгоритмов, например, по критериям: минимизации *СлЗП*, минимаксимизации отдельных свойств и параметров исходного текста и др.

Цель предредактирования – снизить сложность переводческой задачи $СлЗП_{txtiTXT}$ до низкой. На начальном этапе предредактирования происходит отбор фрагментов текста, которым соответствуют высокие значения $СлЗП_i$, то есть

$$txt_i^* \in txt_{iTXT} \mid СлЗП_i > СлЗП_{доп} \quad (42)$$

где $СлЗП_{доп}$ – допустимое значение сложности задачи перевода, т.е. низкое.

Для тех фрагментов текста, для которых сложность задачи перевода является средней или высокой, определяем элементы вектора $C_{оцi}$, значения которых повышают сложность переводческой задачи данного фрагмента: $C_{оцi}|СлЗП_i > СлЗП_{доп}$.

Пусть множество U_{AOT} – универсальное множество всех возможных алгоритмов обработки текста. *Іоц* – множество индексов-количество оценок свойств и параметров текстов. При этом множества *ОЦ* и *Іоц* являются биективными, т.е. *Іоц* ↔ *ОЦ*. Каждому элементу *іоц* ∈ *Іоц* однозначно сопоставлено подмножество алгоритмов оптимизации $AO_{іоц} \subseteq U_{AOT}$. Тогда $СмАО = (AO_{іоц})_{іоц \in Іоц}$ — счетное семейство множеств алгоритмов предредактирования с целью оптимизации оценок текста.

$$СмАО = (AO_{іоц})_{іоц \in Іоц} = \bigcup_{ice \in Ice} AO_{іоц} = \{x: \exists іоц \ x \in AO_{іоц}\} \quad (43)$$

Для каждого найденного элемента матрицы $C_{оцi}$, соответствующего условию $C_{оцi}|СлЗП_i > СлЗП_{доп}$, в зависимости от его значения и ограничений алгоритмов предредактирования определяется стратегия предредактирования (таблица 2.2). Выбор алгоритма зависит от критерия оптимизации выбранного значения.

Таблица 2.2 – Стратегии предредактирования

Значение параметра	**Действие**
В допустимых пределах	Автоматическое предредактирование при помощи алгоритмов
Выходит за допустимые пределы	Полуавтоматическое предредактирование текста с привлечением пользователя

Далее производим редактирование текста в соответствии с имеющимися методами и алгоритмами предредактирования, получаем текст txt'_{iTXT} и переходим к этапу перевода. Формально функцию предредактирования $F_{предред}$ можно представить следующим образом:

$$F_{педред} : txt_{iTXT,} СлЗП,\ C_{оц},\ W_{оц} \rightarrow txt'_{iTXT} \mid (СлЗП \rightarrow min, СмАО,\ \overrightarrow{Кпер_{iПЕР}},\ \overrightarrow{Спер_{iПЕР,дп_{iДП}}}). \quad (44)$$

Схематично, функция предредактирования исходного текста представлена на рисунке 2.2.

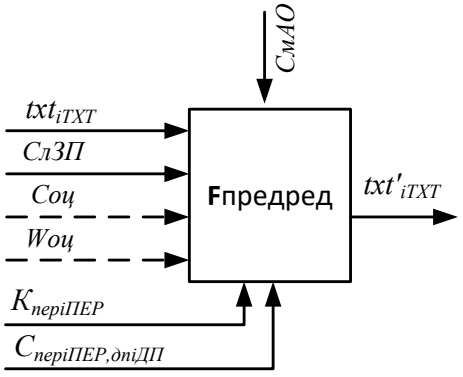

Рисунок 2.2 – Предредактирование исходного текста

2.1.4 Перевод

Цель данного этапа – получение качественного перевода текста txt'_{iTXT} с языка $яз_{вх}$ для перевода на язык $яз_{вых}$. Будем считать, что **качественный перевод** – передача смысла текста на языке оригинала в соответствии с языковыми и культурными традициями языка перевода согласно требованиями конечного реципиента.

Этап перевода включает следующие шаги:

Шаг 1. Сегментация текста на N отрезков. Сегментация текста – процесс разделения письменного текста на значимые отрезки текста, достаточные для передачи его смыслового содержания без искажения общего смысла текста или высказывания.

Далее для каждого i-го отрезка текста при $i=\overline{1,N}$:

Шаг 2. Анализ смысла исходного текста на языке $яз_{вх,}$ на основании компетенции переводчика относительно языка $яз_{вх}$ ($\overrightarrow{Кпер_{iПЕР}} \mid яз_{вх}$) и специализации переводчика в предметной области исходного текста $дп_{iДП}$ $(\overrightarrow{Спер_{iПЕР,}дп_{iДП}})$ с целью определить набор семантических единиц, которые он описывает, то есть его смысл:

$$f_a : (txt'_{iTXT}, \overrightarrow{Кпер_{iПЕР}} \mid яз_{вх}, \overrightarrow{Спер_{iПЕР,}дп_{iДП}}) \rightarrow СМ_i. \quad (45)$$

Шаг 3. Генерация текста, который должен максимально воссоздавать смысл исходного текста $СМ_i$ согласно требованиям $ТР|txt_{iTXT}$, но уже средствами языка $яз_{вых}$ на основании компетенции переводчика относительно языка $яз_{вых}$ ($\overrightarrow{Кпер_{iПЕР}} \mid яз_{вых}$) и специализации переводчика в предметной области исходного текста $дп_{iДП}$ $(\overrightarrow{Спер_{iПЕР,}дп_{iДП}})$:

$$f_g : (СМ_i, \overrightarrow{Кпер_{iПЕР}} \mid яз_{вых}, \overrightarrow{Спер_{iПЕР,}дп_{iДП}}) \rightarrow txt'_{iTXT} \mid (СМ_j \rightarrow СМ_i, ТР|txt_{iTXT}). \quad (46)$$

Формально обобщенную функцию перевода $F_{пер}$ можно представить следующим образом:

$$F_{пер} : (txt'_{iTXT}, \overrightarrow{Кпер_{iПЕР}}, \overrightarrow{Спер_{iПЕР,}дп_{iДП}}) \rightarrow txt_{jTXT} \mid (СМ_j \rightarrow СМ_i, ТР|txt_{iTXT}). \quad (47)$$

Схематично, функция перевода представлена на рисунке 2.3.

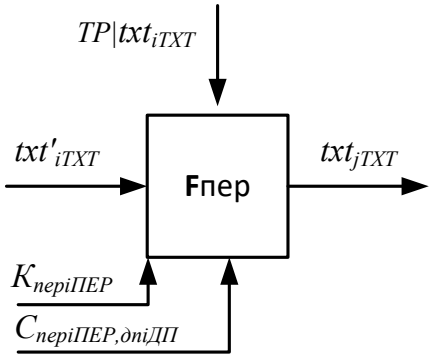

Рисунок 2.3 – Перевод текста

2.1.5 Оценка качества перевода

Проверка качества перевода – комплексная задача оценки оптимальности передачи смысла исходного текста txt_{iTXT} на языке $яз_{вх}$ средствами $яз_{вых}$ в виде текста txt_{jTXT}. Критериями оптимальности являются критерии качества перевода:

$$кк \in КК: КК = \{кк_0, кк_1, \ldots, кк_{nКК}\}. \quad (48)$$

На основании требований к переводу $TP|txt_{iTXT}$ и применимости в рамках домена приложения $дn_{iДП}$ формируется множество применимых критериев качества $ПКК = КК|\ TP|txt_{iTXT}, ПКК \subseteq КК$ и множество значений $w_{iпкк}$ значимости для каждого из применимых критериев, которые формируют нормированную матрицу значимости критериев качества $\overline{W_{ПКК}}$ размерностью $1 \times M$, где M – общее число применимых критериев.

Проверка выполняется контролером качества перевода, в качестве которого выступает некоторый $пер_{jПЕР}$, обладающий соответствующей компетенцией $\overrightarrow{Кпер_{jПЕР}}$ и специализацией $\overrightarrow{Спер_{jПЕР},дn_{iДП}}$. Переводчик, переводивший текст может выступать контролером качества перевода, в случае если выполняется самопроверка перевода, тогда считается, что $пер_{jПЕР} = пер_{iПЕР}$, но следует учитывать, что в таком случае на оценку качества перевода будут влиять значения компетенции $\overrightarrow{К'пер_{iПЕР}}$ и специализации $\overrightarrow{С'пер_{iПЕР},дn_{iДП}}$, так как в процессе выполнения перевода происходит дообучение переводчика и значения его компетенции и специализации изменяются.

Далее для каждой *i-й* пары однозначно сопоставленных друг другу фрагментов текстов txt_{iTXT} *и* txt_{jTXT} при $i=\overline{1,N}$ производится оценка оптимальности согласно каждому *k-му* применимому критерию качества при $k=\overline{1,M}$ на основании $СМ_i$ и свойств и параметров текста $СВ|txt_{jTXT}$ и $ГП|txt_{jTXT}$. Точность определения значения оптимальности зависит от компетенции и специализации контролера качества. Оценка оптимальности перевода по *k-му* критерию *i-го* фрагмента текста:

$$f_{коо} : ((СМ_i)_i, (СВ|txt_{jTXT}, ГП|txt_{jTXT})_i) \rightarrow ОО_{i,k} \mid (ПКК_k, \overrightarrow{Кпер_{jПЕР}}, \overrightarrow{Спер_{jПЕР},дn_{iДП}}) \quad (49)$$

Оценка качества перевода *i-го* фрагмента текста – это определитель матрицы, полученной в результате перемножения матрицы оценок оптимальности фрагмента переведенного текста $ОО_i$ по критериям $ПКК$ и транспонированной матрицы значимости применимых критериев качества $\overline{W^T_{ПКК}}$:

$$ОК_i = \left| ОО_i \times \overline{W^T_{ПКК}} \right| \quad (50)$$

Результирующая оценка качества перевода текста txt_{iTXT} на язык $яз_{вых}$ – отношение суммы оценок качества перевода отдельных фрагментов текста к их количеству, то есть

$$ОК = \frac{\sum_N^i ОК_i}{N} \quad (51)$$

Оценка качества перевода позволяет оценить эффективность работы переводчика и определить стратегию обработки переведенного текста, в случае если *ОК* вне допустимых значений. Возможные варианты постобработки текста приведены в таблице 2.3.

Таблица 2.3 – Действия в зависимости от оценки качества перевода

Оценка качества	**Действие**
Крайне низкая	Отказ от перевода, перевод другим переводчиком
Низкая	Постредактирование текста с привлечением эксперта
Средняя	Самостоятельное / автоматическое постредактирование текста
Высокая	–

Диапазоны значений *ОК,* соответствующие крайне низкому, низкому, среднему и высокому уровням определяются на основе анализа применимых критериев качества перевода, их важности, а также целевых значений согласно требованиям к переводу $TP|txt_{iTXT}$.

Формально функцию оценивания качества перевода F_{kk} можно представить следующим образом:

$$F_{kk} : (txt_{iTXT}, txt_{jTXT}) \rightarrow ОК|(TP|txt_{iTXT}, \overrightarrow{Кпер_{jПЕР}}, \overrightarrow{Спер_{jПЕР,дп_{iДП}}}). \quad (52)$$

Схематично, функция оценки качества перевода представлена на рисунке 2.4.

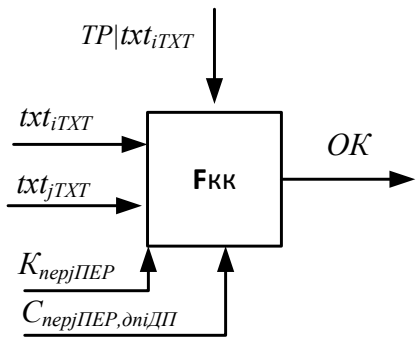

Рисунок 2.4 – Оценка качества перевода

2.1.6 Постредактирование переведенного текста

Постредактирование – это процесс автоматического/полуавтоматического изменения свойств/параметров переведенного текста с целью оптимизации оценки качества перевода. Постредактирование выполняется, если оценка качества выполненного перевода ниже допустимого либо целевого значения $ОК_{TARGET}$.

Цель постредактирования – оптимизировать оценку качества перевода по критерию максимизации, то есть

$$ОК \rightarrow max || ОК_{TARGET}. \quad (53)$$

Для тех фрагментов текста, для которых оценка качества $ОК_i$ ниже высокой либо целевой $ОК$, определяем элементы матрицы $ОО_i$, значения которых понижают качество перевода данного фрагмента:

$$ОО_i | ОК_i < max || ОК_{TARGET}. \quad (54)$$

Пусть множество $U_{АОТ}$ – универсальное множество всех возможных алгоритмов обработки текста. $Iкк$ – множество индексов-количество критериев качества. При этом множества $КК$ и $Iкк$ являются биективными, т.е. $Iкк \leftrightarrow КК$. Каждому элементу $iкк \in Iкк$ однозначно сопоставлено подмножество алгоритмов оптимизации свойств и параметров текста по соответствующему критерию $КК: АОК_{iкк} \subseteq U_{АОТ}$. Тогда $СмАОК = (АОК_{iкк})_{iкк \in Iкк}$ – счетное семейство множеств алгоритмов постредактирования с целью оптимизации свойств и параметров текста по критериям качества.

$$СмАОК = (АОК_{iкк})_{iкк \in Iкк} = \bigcup_{iкк \in Iкк} АОК_{iкк} = \{x: \exists iкк \; x \in АОК_{iкк}\} \quad (55)$$

Для каждого найденного элемента матрицы $ОО_i$, соответствующего условию (54), в зависимости от его значения и ограничений алгоритмов постредактирования определяется стратегия постредактирования (таблица 2.4). Выбор алгоритма зависит от критерия оптимизации выбранного значения.

Таблица 2.4 – Стратегии постредактирования

Значение параметра	**Действие**
В допустимых пределах	Автоматическое постредактирование при помощи алгоритмов
Выходит за допустимые пределы	Полуавтоматическое постредактирование текста с привлечением пользователя

Далее каждый найденный фрагмент редактируется при помощи алгоритмов постредактирования до тех пор, пока не будут достигнуты значения OO_i, при которых выполняется условие (54). Повторная оценка OO_i осуществляется по формуле (49). На выходе получаем текст $txt_{итгTXT}$. Эффективность применения алгоритмов постредактирования зависит от специализации и компетенции переводчика, осуществляющего постредактирование $\overrightarrow{Кпер_{jПЕР}}$, $\overrightarrow{Спер_{jПЕР,дп_{iДП}}}$.

Формально функцию постредактирования $F_{постред}$ можно представить следующим образом:

$$F_{постред} : (txt_{iTXT}, txt_{jTXT}, ОК, ОО, \overline{W_{ПКК}}) \rightarrow txt_{итгTXT} \mid$$
$$(ОК \rightarrow max \mid\mid ОК_{TARGET}, СмАОК, \overrightarrow{Кпер_{iПЕР}}, \overrightarrow{Спер_{iПЕР,дп_{iДП}}}). \quad (56)$$

Схематично, функция постредактирования представлена на рисунке 2.5.

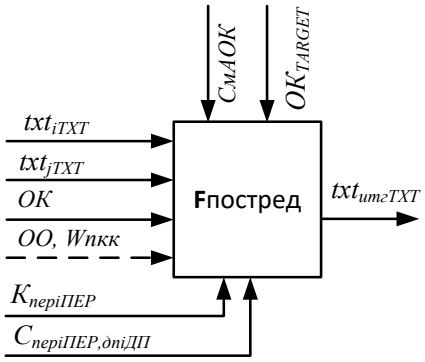

Рисунок 2.5 – Постредактирование переведенного текста

2.1.7 Обучение переводчика

Обучение переводчика – процесс формирования и/или обновления компетенций и специализаций переводчика на основе изучения обучающих текстов и/или полученного практического опыта перевода. Важно отметить, что процесс дообучения происходит непрерывно в процессе работы.

Каждый переводчик обладает множеством компетенций и специализаций в зависимости от известных переводчику языков $ЯЗ|пер_{iПЕР}$ и доменов приложения $ДП|пер_{iПЕР}$.. Множество компетенций переводчика:

$$кп \in К: К = \{кп_0, кп_1, \ldots кп_{nK}\}, \quad (57)$$

где $кп_{iКП}$ определяется по формуле (36) для некоторой языковой пары $(яз_{вх}, яз_{вых})$.

Множество специализаций переводчика:

$$cn \in C:\ C = \{cn_0, cn_1, \dots\ cn_{nC}\}, \tag{58}$$

где $cn_{iСП}$ определяется по формуле (35) для некоторого домена приложения *дп*.

Обучение проводится на реальных текстах с проверкой и работой над ошибками. В зависимости от состава обучающих текстов переводчик осваивает специализацию по предметной области, например, медицина или технический перевод. Чем больше переводчик учится на специальных текстах, тем выше качество перевода текстов соответствующей предметной области, но и качество перевода других областей растет просто за счет лучшего освоения языка. В процессе обучения переводчик увеличивает объем знаний, развивает умения и научается тому, как используются отдельные слова и их сочетания в разных доменах для выражения семантических единиц.

Множество **обучающих текстов**:

$$om \in OT:\ OT = \{om_0,\ om_1, \dots\ om_{nOT}\} \tag{59}$$

где om_{iOT} – это пара значений $txt_{вх}$ на языке оригинала $яз_{вх}$ *и* $txt_{вых}$ на языке перевода $яз_{вых}$ однозначно сопоставленных друг другу по смыслу (набору выражаемых семантических единиц), принятая за эталон, то есть

$$om_{iOT} = \{(txt_{вх};\ txt_{вых}):\ CM|\ txt_{вх} = CM \mid txt_{вых}\}. \tag{60}$$

В общем виде функцию обучения переводчика можно записать следующим образом:

$$F_{об} : (OT,\ K_{о,періПЕР},\ C_{о,періПЕР}) \rightarrow K_{періПЕР},\ C_{періПЕР.} \tag{61}$$

Схематично, функция обучения переводчика на рисунке 2.6.

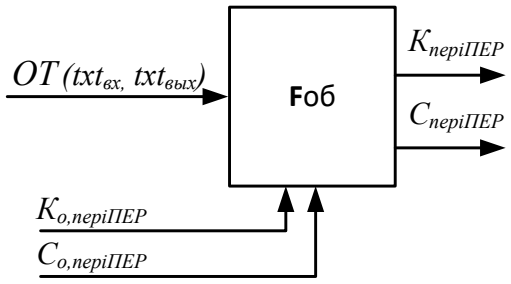

Рисунок 2.6 – Обучение переводчика

2.1.8 Обобщенная модель перевода

Обобщенно, разработанная модель процесса перевода представлена на рисунке 2.7.

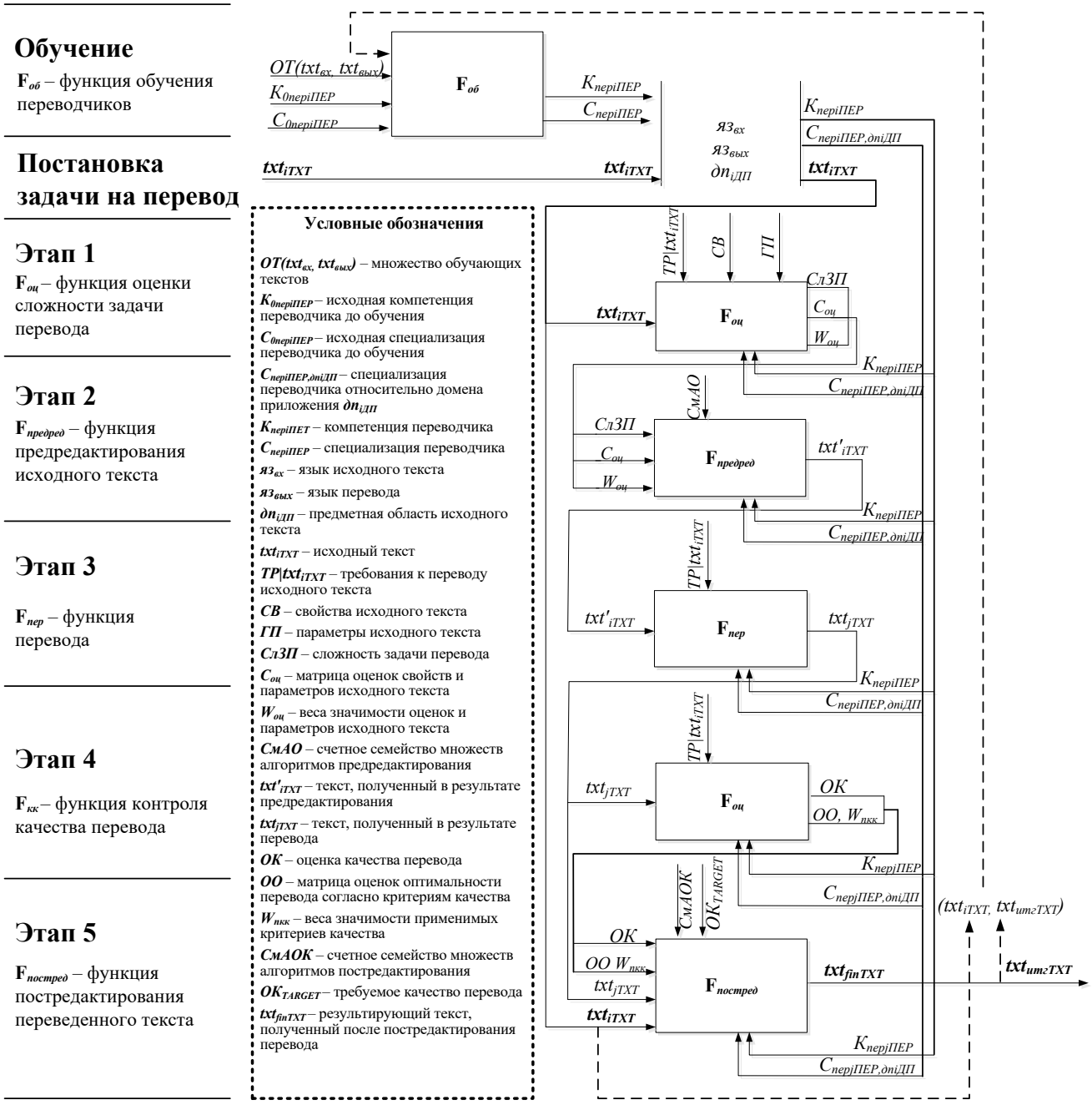

Рисунок 2.7 – Модель процесса перевода

Результаты выполненного моделирования показывают, что уже на этапе оценки исходного текста, возможно предсказать ожидаемое качество перевода на основе параметров исходного текста и компетенции и специализации переводчика [101].

Дальнейшее исследование предполагает более глубокую декомпозицию и моделирование, с четом того, что в современных системах МП эти этапы, как правило, не реализованы либо реализованы частично. Так, в системах МП не реализован этап переводческого процесса, который выполняется при «ручном переводе», а именно оценка сложности задачи перевода. На этом этапе переводчик оценивает вероятность получения качественного перевода, то есть соответствующего требованиям заказчика, и, если эта вероятность низкая, выбирает стратегию предредактирования исходного текста с целью повышения вероятности получения качественного перевода.

В соответствии с методологией определения сложности задачи перевода, описанной в п.2.1.2, для расчета сложности задачи перевода необходимо рассчитать веса значимости оценок текста $W_{оц} = \{w_{1оц}, w_{2оц}, \ldots w_{коц}\}$ для текстов на языке $яз_{вх}$ при переводе на язык $яз_{вых}$. Опишем модель решения задачи поиска весов значимости признаков текста для оценки сложности задачи перевода при условии наличия некоторой экспериментальной выборки, позволяющей провести обучение модели и сравнение полученных теоретических результатов с экспериментальными.

2.2 Математическая модель поиска весов значимости признаков текста

Для поиска весов значимости признаков текста воспользуемся численным методом наименьших квадратов [102], при котором минимизируется сумма квадратов отклонений эмпирических (фактических) значений результативного признака от теоретических, полученных по уравнению (37), т.е.

$$S(w)= \sum_{i=1}^{R} (OK_i-\widehat{OK}_i(C_{оцi}, w))^2,$$

$$S(w) = \sum_{i=1}^{R} (OK_i-w_0+w_{1оц}C_{оцi_1}+w_{2оц}C_{оцi_2}+ \cdots+w_{оцk}C_{оцik})^2 \rightarrow min \ , \qquad (62)$$

где R – объем экспериментальной выборки.

Для решения задачи минимизации необходимо найти стационарные точки функции $S(w)$, продифференцировав её по искомым параметрам w и приравняв производные к нулю

$$\sum_{i=1}^{R} (OK_i - \widehat{OK}_i(C_{оцi}, w)) \frac{\partial \widehat{OK}_r(C_{оцi}, w)}{\partial w} = 0 . \quad (63)$$

Получаем систему k нормальных уравнений с k неизвестными:

$$\begin{cases} \sum OK = Rw_0 + w_{1оц} \sum C_{оц1} + w_{2оц} \sum C_{оц2} + \ldots + w_{kоц} \sum C_{оцk} \\ \sum OK \cdot C_{оц1} = w_0 \sum C_{оц1} + w_{1оц} \sum C_{оц1}^2 + w_{2оц} \sum C_{оц2} C_{оц1} + \ldots + w_{kоц} \sum C_{оцk} C_{оц1} \\ \ldots \\ \sum OK \cdot C_{оцk} = w_0 \sum C_{оцk} + w_{1оц} \sum C_{оц1} C_{оцk} + w_{2оц} \sum C_{оц2} C_{оцk} + \ldots + w_{kоц} \sum C_{оцk}^2 \end{cases}$$

Решение этой системы уравнений дает нам общую формулу поиска весов значимости $W_{оц}$ в матричной форме:

$$W_{оц} = (C_{оц}^T \cdot C_{оц})^{-1} \cdot C_{оц}^T \cdot OK = (\frac{1}{R} C_{оц}^T \cdot C_{оц})^{-1} \frac{1}{R} C_{оц}^T \cdot OK. \quad (64)$$

Предложенный алгоритм позволяет расширить область применения метода наименьших квадратов для поиска весов значимости параметров исходного текста для вероятностной оценки ожидаемого качества его перевода на целевой язык.

2.3 Математическая постановка задачи машинного перевода

Для формализации задачи машинного перевода воспользуемся методом максимального правдоподобия, который наиболее часто используется в задачах машинного обучения из-за высокой сходимости и вычислительной эффективности [103]. Метод максимального правдоподобия – статистический метод, используемый для оценки параметров распределения на основе наблюдаемых данных. Суть метода заключается в выборе таких значений параметров, которые максимизируют вероятность (или правдоподобие) получения наблюдаемых данных. Стоит отметить, что метод применим не только при традиционном выборочном подходе, но и при использовании байесовского подхода к анализу данных. Байесовские оценки обычно эквивалентны оценкам максимального правдоподобия с ограничениями [104].

Способ измерения близости к истинному параметру – вычисление квадратичной разницы между оценочными и истинными значениями параметров, где математическое ожидание вычисляется над L обучающими выборками из данных,

генерирующих распределение. Эта параметрическая среднеквадратичная ошибка уменьшается с увеличением L, и для больших L нижняя граница неравенства Крамера-Рао показывает, что среди прочих сходящихся функций у оценки максимального правдоподобия среднеквадратичная ошибка минимальна и по мере того, как число обучающих выборок приближается к бесконечности, оценка максимального правдоподобия сходится к истинному значению параметра [105].

В контексте задач МП, метод максимального правдоподобия может быть применен для оценки параметров модели языка или модели перевода. Он позволяет найти такие значения параметров, которые максимизируют вероятность генерации (или перевода) целевых предложений на основе имеющихся параллельных корпусов или наборов предложений.

Математически, для систем МП задачу перевода можно формализовать через метод максимизации функции правдоподобия следующим образом.

Опишем условия задачи, пусть:

1) txt_{iTXT} – исходный текст на языке $яз_{вх}$;

2) txt_{jTXT} – переведенный текст на языке $яз_{вых}$;

3) TXT_{itrg} – множество всех возможных вариантов перевода текста txt_{iTXT} на язык $яз_{вых}$:

$$TXT_{itrg} = \{txt_0, txt_1, \ldots, txt_j\},$$

где j – обще число вариантов перевода, $txt_{jTXT} \in TXT_{itrg}$;

4) OK_{itrg} – множество нормированных оценок качества перевода текста txt_{iTXT} в соответствии с требованиями к переводу $TP|txt_{iTXT}$ *для* всех возможных вариантов перевода TXT_{itrg}:

$$OK_{itrg} = \{OK_0, OK_1, \ldots, OK_j\},$$

где OK_j – оценка качества для j-*го* варианта переведенного текста.

5) Каждому варианту перевода соответствует одна оценка качества перевода, то есть множества TXT_{itrg} и OK_{itrg} биективны: $TXT_{itrg} \leftrightarrow OK_{itrg}$;

6) $[minOK; maxOK]$ – диапазон значений оценок качества перевода OK_{itrg};

7) $OK_{доп}$ – минимально допустимое значение критерия «Высокая оценка качества перевода» при допущении, что чем выше значение OK_j, тем лучше;

9) $КП_i$ – нечёткое подмножество множества $ОК_{itrg}$, определяющее принадлежность элементов множества $ОК_{itrg}$ классу «Высокая оценка качества перевода текста txt_{iTXT}»:

$$КП_i = \{(ОК, \mu_{КПi}(ОК)) | ОК \in ОК_{itrg}\};$$

10) $\mu_{КПi}(ОК)$ – функция принадлежности, указывающая в какой степени текст txt с оценкой $ОК$ принадлежит нечеткому множеству $КП_i$;

11) $\mu_{КПi}(ОК) \in [0; 1]$ и имеет вид логистической кривой (рисунок 2.8):

$$\mu_{КПi}(ОК) = \frac{1}{1 + e^{-\left(\frac{ОК - ОК_{доп}}{maxОК - ОК_{доп}}\right)2\pi}} \quad (65)$$

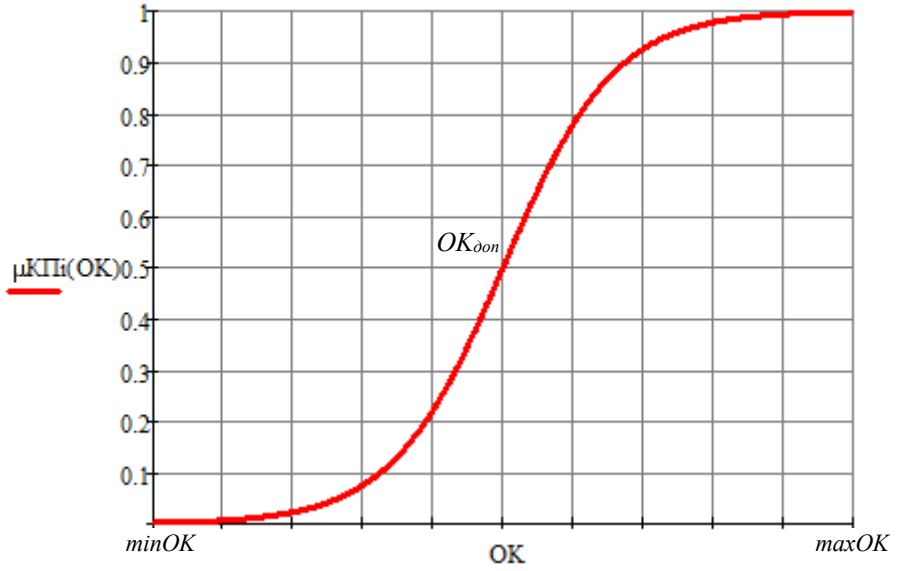

Рисунок 2.8 – График функции принадлежности $\mu_{КПi}(ОК)$

Требуется максимизировать правдоподобие сгенерированного системой МП текста txt_{jTXT}, то есть вероятность того, что txt_{jTXT} примет такое значение, при котором $\mu_{КПi}(ОК)$ будет максимальна. Тогда логарифмическая функция правдоподобия машинного перевода $F_{МП}$ примет вид:

$$F_{МП}(\theta, \mu_{КПi}(ОК)) = lnP_{\theta}(max\ \mu_{КПi}(ОК)) \rightarrow \max_{\theta}, \quad (66)$$

где θ – параметры системы МП из множества исполнителей перевода, или переводчиков: $пер_{iПЕР} \in ПЕР$, максимизирующие вероятность P получить максимальное значение функции принадлежности $\mu_{КПi}(ОК)$.

Решение поставленной задачи лежит в области оптимизации и совершенствования алгоритмов генерации переведенного текста. Однако, в рамках настоящего исследования стоит цель повышения качества переведенного текста на этапе подготовки к переводу до непосредственной генерации переведенного текста, поэтому далее рассмотрим задачу автоматического оптимизационного предредактирования.

2.4 Математическая постановка задачи оптимизационного предредактирования

Задача оптимизационного предредактирования состоит в том, чтобы максимизировать правдоподобие, то есть вероятность того, что при параметрах Ψ предредактора, текст txt'_{iTXT} на языке $яз_{вх}$ будет эквивалентен txt_{iTXT} по смыслу, понятен системе МП $пер_{iПЕР}$ и оценка качества $ОК_j$ перевода txt_{jTXT} относительно txt_{iTXT} при генерации перевода из txt'_{iTXT} будет максимальной. Далее опишем задачу более подробно.

Условия задачи выглядят в следующем виде, пусть:

1) txt'_{iTXT} – текст на языке $яз_{вх}$, созданный системой автоматического оптимизационного предредактирования, такой, при котором $СМ'_i \rightarrow СМ_i$ и $txt'_{iTXT} \neq txt_{iTXT}$, где $СМ$ – смысл или упорядоченный набор семантических единиц, описываемый текстом: $СМ = \{(се_0, се_1, \ldots, се_{псм}): се_{iсм} \in СЕ\}$, $СМ'_i$ и $СМ_i$ – смыслы txt'_{iTXT} и txt_{iTXT}, соответственно.

2) $ТХТ_{isrc}$ – множество всех возможных вариантов предредактированного текста, т.е. выражения смысла $СМ_i$ текста txt_{iTXT} на языке $яз_{вх}$:

$$ТХТ_{isrc} = \{txt_0, txt_1, \ldots, txt_k\},$$

где k – обще число вариантов предредактированного текста, причем txt_{iTXT}, $txt'_{iTXT} \in ТХТ_{isrc}$;

3) $мСлЗП_{isrc}$ – множество оценок сложности задачи перевода вариантов предредактирования текста txt_{iTXT} в соответствии с компетентностью $\overrightarrow{Кпер_{iПЕР}}$ и специализацией $\overrightarrow{Спер_{iПЕР},дп_{iДП}}$ системы МП $пер_{iПЕР}$ для всех возможных вариантов предредактированного текста $ТХТ_{isrc}$:

$$мСлЗП_{isrc} = \{СлЗП_0, СлЗП_1, \ldots, СлЗП_k\};$$

4) Каждому варианту предредактированного текста соответствует одна оценка сложности задачи перевода для системы МП $пер_{iПЕР}$, то есть множества TXT_{isrc} и $мСлЗП_{isrc}$ биективны: $TXT_{isrc} \leftrightarrow мСлЗП_{isrc}$;

5) $СлЗП_k \in мСлЗП_{isrc}$ – оценка сложности задачи перевода варианта предредактированного текста txt'_{iTXT} для системы МП $пер_{iПЕР}$;

6) [$minСлЗП$; $maxСлЗП$] – диапазон значений оценок сложности задачи перевода $мСлЗП_{isrc}$;

7) $СлЗП_{доп}$ – максимально допустимое значение критерия «Низкая сложность задачи перевода» при допущении, что чем ниже значение $СлЗП_k$, тем лучше;

8) $нСлЗП_i$ – нечёткое подмножество множества $мСлЗП_{isrc}$, определяющее принадлежность элементов множества $мСлЗП_{isrc}$ и соответствующих элементов множества TXT_{isrc} классу «Низкая сложность задачи перевода»:

$$нСлЗП_i = \{(СлЗП, \mu_{нСлЗПi}(СлЗП))| \; СлЗП \in мСлЗП_{isrc}\};$$

9) $\mu_{нСлЗПi}(СлЗП)$ – функция принадлежности, указывающая в какой степени текст txt с оценкой $СлЗП$ принадлежит нечеткому множеству $нСлЗП_i$;

10) $\mu_{нСлЗПi}(СлЗП) \in [0; 1]$ и имеет вид логистической кривой (рисунок 2.9):

$$\mu_{нСлЗПi}(СлЗП) = \frac{1}{1+e^{\left(\frac{СлЗП-СлЗП_{доп}}{|minСлЗП-СлЗП_{доп}|}\right)2\pi}} \quad (67)$$

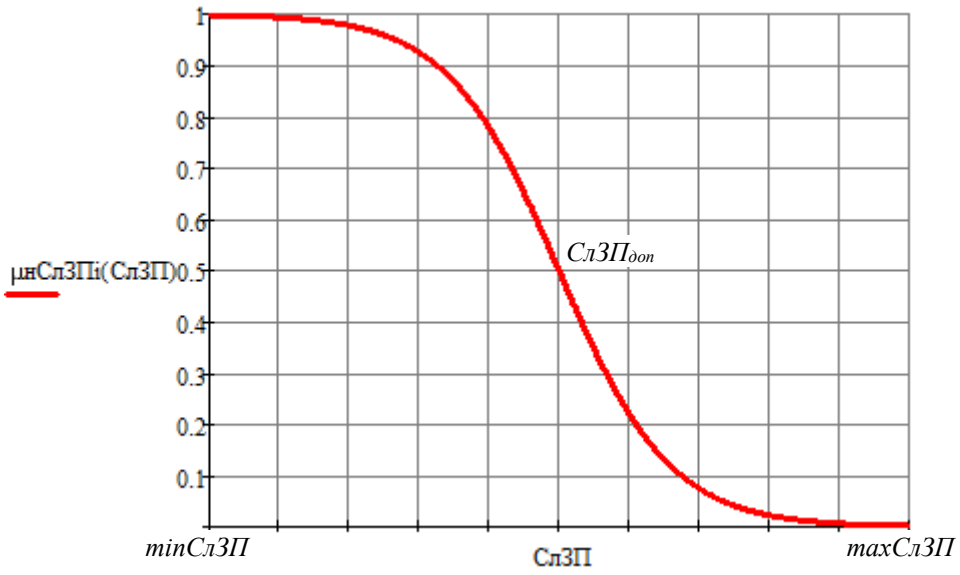

Рисунок 2.9 – График функции принадлежности $\mu_{нСлЗПi}(СлЗП)$

Требуется максимизировать правдоподобие сгенерированного системой оптимизационного предредактора текста txt'_{iTXT}, то есть вероятность того, что txt'_{iTXT} примет такое значение, при котором $\mu_{нСлЗПi}(СлЗП)$ будет максимальна.

В дискретном случае функция правдоподобия $F_{АОПР}(\Psi, \mu_{нСлЗПi}(СлЗП))$ – вероятность выборке $\mu_{нСлЗПi}(СлЗП) = \{\mu_0, \mu_1, \dots, \mu_l\}$ в рассматриваемой серии экспериментов равняться $\{max\ \mu_{нСлЗПi}(СлЗП)_0, max\ \mu_{нСлЗПi}(СлЗП)_1, \dots, max\ \mu_{нСлЗПi}(СлЗП)_l\}$. Эта вероятность меняется в зависимости от Ψ:

$$F_{АОПР}(\Psi, \mu_{нСлЗПi}(СлЗП)) = \prod_{l=1}^{L} F_{АОПР}(\mu_{нСлЗПi}(СлЗП)_l) =$$
$$P_\Psi(\mu_0 = max\ \mu_{нСлЗПi}(СлЗП)_0) \cdot \dots \cdot P_\Psi(\mu_l = max\ \mu_{нСлЗПi}(СлЗП)_l) = \quad (68)$$
$$P_\Psi(\mu_0 = max\ \mu_{нСлЗПi}(СлЗП)_0), \dots, \mu_l = max\ \mu_{нСлЗПi}(СлЗП)_l),$$

где l – номер семпла в обучающей выборке объемом L.

Тогда логарифмическая функция правдоподобия автоматического оптимизационного предредактирования $F_{АОПР}$ имеет вид:

$$L_{АОПР}(\Psi, \mu_{нСлЗПi}(СлЗП)) = lnP\ \Psi(max\ \mu_{нСлЗПi}(СлЗП)), \quad (69)$$

где Ψ – параметры системы автоматического оптимизационного предредактирования, максимизирующие вероятность P получить максимальное значение функции принадлежности $\mu_{нСлЗПi}(СлЗП)$.

Поскольку $ln(y)$ монотонна, то точки максимума $F_{АОПР}(\Psi, \mu_{нСлЗПi}(СлЗП))$ и $L_{АОПР}(\Psi, \mu_{нСлЗПi}(СлЗП))$ совпадают, и оценкой максимального правдоподобия можно назвать точку максимума функции $L_{АОПР}(\Psi, \mu_{нСлЗПi}(СлЗП))$ по Ψ. Задача оптимизации, таким образом, заключается в поиске оценки максимального правдоподобия $\widehat{\Psi}$ вектора параметров Ψ, или:

$$\widehat{\Psi} = arg \max_{\Psi} L_{АОПР}(\Psi, \mu_{нСлЗПi}(СлЗП)) \quad (70)$$

Далее рассмотрим решение поставленной задачи оптимизации методом градиентного спуска.

2.5 Метод градиентного спуска для решения задачи автоматического оптимизационного предредактирования

Для решения задачи автоматического оптимизационного предредактирования воспользуемся методом градиентного спуска (подъема) [106]. Метод градиентного спуска и является универсальным методом решения задачи оптимизации и широко используется в задачах машинного обучения на больших объемах данных ввиду высокой скорости сходимости, гибкости в выборе шага, точности и критерия остановки, а также возможности обновлять параметры модели на отдельных подмножествах обучающей выборки данных (семпле), что делает его вычислительно эффективным.

Для решения задачи автоматического оптимизационного предредактирования необходимо найти градиент логарифмической функции правдоподобия $L_{АОПР}(\Psi, \mu_{нСлЗПi}(СлЗП))$ – вектор, который показывает направление возрастания функции.

Учитывая, что Ψ – вектор параметров системы автоматического оптимизационного предредактирования и $\Psi = \{\psi_1, \psi_2, \ldots, \psi_m\}$, где m – количество параметров модели, градиент функции $L_{АОПР}(\Psi, \mu_{нСлЗПi}(СлЗП))$ может быть найден по формуле:

$$\nabla L_{АОПР}(\Psi) = (\partial L_{АОПР}/\partial \psi_1, \partial L_{АОПР}/\partial \psi_2, \ldots, \partial L_{АОПР}/\partial \psi_m), \quad (71)$$

где $\partial L_{АОПР}/\partial \psi_m$ – частная производная функции правдоподобия по m-ному параметру.

Обновление параметров Ψ происходит итеративно для каждого $\psi_m \in \Psi$:

$$\Psi^{[s+1]} = \Psi^{[s]} + \alpha \cdot \nabla L_{АОПР}(\Psi^{[s]}), \quad (72)$$

где s – шаг оптимизации, $s \in [0;S]$ и S – общее число шагов оптимизации, а $\Psi^{[0]}$ – начальное приближение параметров модели; α – скорость обучения, т.е. положительное число, определяющее размер шага на каждой итерации.

Для оценки сходимости используется евклидова норма градиента функции $\nabla L_{АОПР}(\Psi)$:

$$\| \nabla L_{АОПР}(\Psi) \| = \sqrt{\left(\frac{\partial L_{АОПР}}{\partial \psi_1}\right)^2 + \left(\frac{\partial L_{АОПР}}{\partial \psi_2}\right)^2 + \ldots + \left(\frac{\partial L_{АОПР}}{\partial \psi_m}\right)^2}. \quad (73)$$

Уменьшение нормы градиента указывает на сходимость оптимизации. Если норма градиента не снижается, это свидетельствует о медленной сходимости и необходимости изменения параметров оптимизации, например, скорости обучения α.

Оптимизация выполняется, пока норма градиента не достигла заданной точности ε, критерий остановки:

$$\left\| \nabla L_{АОПР}\left(\Psi^{[s]}\right) \right\| \leq \varepsilon. \quad (74)$$

Предложенный алгоритм позволяет расширить область применения оптимизационного метода градиентного спуска путём использования элементов нечеткой логики в выражении функции правдоподобия через функцию принадлежности полученного текста низкой сложности задачи перевода для выбранной системы МП.

Выводы по второй главе

В ходе математического и процессного моделирования работы переводчика впервые обоснована целесообразность и разработана методология оценки сложности переводческой задачи. Результаты выполненного моделирования показывают, что уже на этапе оценки исходного текста, возможно предсказать ожидаемое качество перевода на основе параметров исходного текста и компетенции, и специализации переводчика.

Впервые разработана математическая модель расчета сложности задачи перевода и предложено использовать данную оценку в качестве критерия оптимизации при решении задачи предредактирования исходного текста на этапе подготовки к машинному переводу. Предложенное решение задачи оптимизационного предредактирования расширяет область применения метода градиентного спуска (подъема) за счет применения элементов нечеткой логики при формализации функции потерь.

Далее опишем методологию реализации разработанных моделей для решения задачи оптимизационного предредактирования русскоязычных текстов с целью повышения качества их перевода на английский язык.

ГЛАВА 3 МЕТОДИКА РЕАЛИЗАЦИИ ОПТИМИЗАЦИОННОГО ПРЕДРЕДАКТИРОВАНИЯ

3.1 Модель оптимизационного предредактирования

При реализации модели оптимизационного предредактирования предлагается использовать опорную языковую модель T5. T5 – это трансформерная архитектура моделей обработки естественного языка, которая использует подход «текст-в-текст» для решения различных задач обработки естественного языка и может быть использована для решения задач, таких как перевод, генерация текста, ответы на вопросы и т.д., путем преобразования исходного текста в целевой текст [107]. Обработка текста в модели T5 происходит в два этапа [108]:

Этап 1. Текст токенизируется и последовательно проходит через стек энкодеров. Энкодер получает на вход список векторов (тензоров), обрабатывает его, передавая векторы в слой самовнимания для нормализации, затем передает данные в нейронную сеть с прямой связью, а далее отправляет их к следующему энкодеру. Нормализация состоит в масштабировании входных значений тензора таким образом, чтобы среднее значение было нулем, а дисперсия равнялась единице.

Этап 2. После прохождения стека энкодеров текст передается в стек декодеров для расшифровки векторных значений обратно в слова. Стек декодеров возвращает вектор чисел, который проходит через полносвязный линейный слой нейронной сети. Выходом линейного слоя являются логит-векторы, которые представляют собой распределенную вероятность выходного слова. После применения функции активации векторы заменяются конкретным словом, а матрицы преобразуются в список выходных слов.

Обучение модели T5 проходит на корпусах параллельных текстов, состоящих из пар-примеров входных и выходных данных. Входные данные – исходный текст, а выходные данные – целевой текст.

Параметрами модели $\Psi = \{\psi_1, \psi_2, \ldots, \psi_m\}$ являются веса нейронов сетей энкодера и декодера. Для обучения модели применяется следующий алгоритм:

1 Инициализация параметров модели случайными значениями.

Далее для каждого семпла обучающей выборки:

2 Кодирование входного текста в последовательность чисел.

3 Кодирование выходного текста в последовательность чисел.

4 Генерация целевого текст, используя веса модели.

5 Вычисление ошибки между предсказанным целевым текстом и фактическим выходным текстом.

6 Обновление весов модели, используя алгоритм, описанный в п.2.5.

3.2 Обучение модели оптимизационного предредактирования

Предредактирование в переводе – это перевод с языка *яз$_{вх}$* на язык *яз$_{вх}$* с целью, во-первых, сделать исходный текст более понятным переводчику для адекватного подбора лексических эквивалентов, а во-вторых, минимизировать риски грамматических и стилистических ошибок. Рассматривая процесс перевода с точки зрения теории переводоведения, можно сказать, что это ряд преобразований (переводческих трансформаций), с помощью которых осуществляется перестроение от единиц оригинала к единицам перевода [109]. При генерировании переведённого текста системы МП стремятся к сохранению структуры исходного текста при условии соблюдения норм языка перевода, т.е. фактически применяют прием переводческих трансформаций – калькирование структуры предложений. Однако калькирование – одна из наиболее простых переводческих трансформаций, использование которой позволяет передать смысл текста, но значительно повышает риск стилистической и грамматической ошибки. Зная об этой особенности, возможно построить предложение на языке *яз$_{вх}$* таким образом, чтобы при калькировании структуры на язык *яз$_{вых}$*, переведенный текст был более стилистически точным.

Таким образом, создав корпус тренировочных текстов в паре *яз$_{вх}$*-*яз$_{вх}$*, возможно настроить языковую модель, которая будет преобразовывать текст на языке *яз$_{вх}$* в текст требуемой структуры для повышения качества перевода.

Для оптимизации временных затрат на подготовку исходных данных для тренировки модели предредактирования текста предлагается методика с

использованием обратного перевода для генерирования эталонного предредактированного текста. Структура параллельного корпуса исходных данных: *RefCor*: *[src_ref; tgt_ref]*, где *src_ref* – это оригинал, т.е. текст на языке $яз_{вх}$, *tgt_ref* – это перевод (текст на языке $яз_{вых}$).

Методика обучения модели оптимизационного предредактирования включает следующие этапы [110]**:**

Этап 1. Генерация корпусов (массивов данных)

1) Настраиваем системы МП *MT:tgt-src*, *MT:src-tgt*.

2) При помощи системы *MT:tgt-src* переводим текст *tgt_ref* на язык $яз_{вх}$, получим массив текстовых данных *pre_src* (массив условно предредактированных текстов).

3) При помощи системы *MT:src-tgt* переводим текст *src_ref* на язык $яз_{вых}$, получим массив текстовых данных *tgt1* (*src_ref* → *tgt1*).

4) При помощи системы *MT: src-tgt* переводим текст *pre_src* на язык $яз_{вых}$, получим массив текстовых данных *tgt2* (*src_ref* → *tgt1*).

5) Оцениваем качество выполненного перевода на язык $яз_{вых}$ *tgt1* и *tgt2* относительно эталона *tgt_ref*, получаем массивы оценок *QC_score(tgt1)* и *QC_score(tgt2)*.

Этап 2. Отбор тренировочных данных и обучение оптимизационного предредактора

6) Для дальнейшей работы отберем тренировочный корпус *TrainCor*, включающий пары src_ref_i и pre_src_i, для которых наблюдается повышение оценки качества перевода на английский язык при применении предредактирования и при условии, что ΔQC_score_i является условно значимой d_{max} для выбранного типа оценки:

$$TrainCor = \{(src\_ref_i;\ pre\_src_i):\ \exists\,(tgt1_i,\ tgt2_i)\ |$$
$$QC\_score(tgt2_i) > QC\_score(tgt1_i)\ \&\ \Delta\ QC\_score_i \geq d_{max}\} \quad (75)$$

7) Настроим языковую модель *LM:src-pre_src* для решения задачи автоматического оптимизационного редактирования текстов на языке $яз_{вх}$, в качестве тренировочного корпуса для обучения модели используем полученный корпус параллельных текстов *TrainCor***.**

Схематично, методика обучения модели оптимизационного предредактирования представлена на рисунке 3.1.

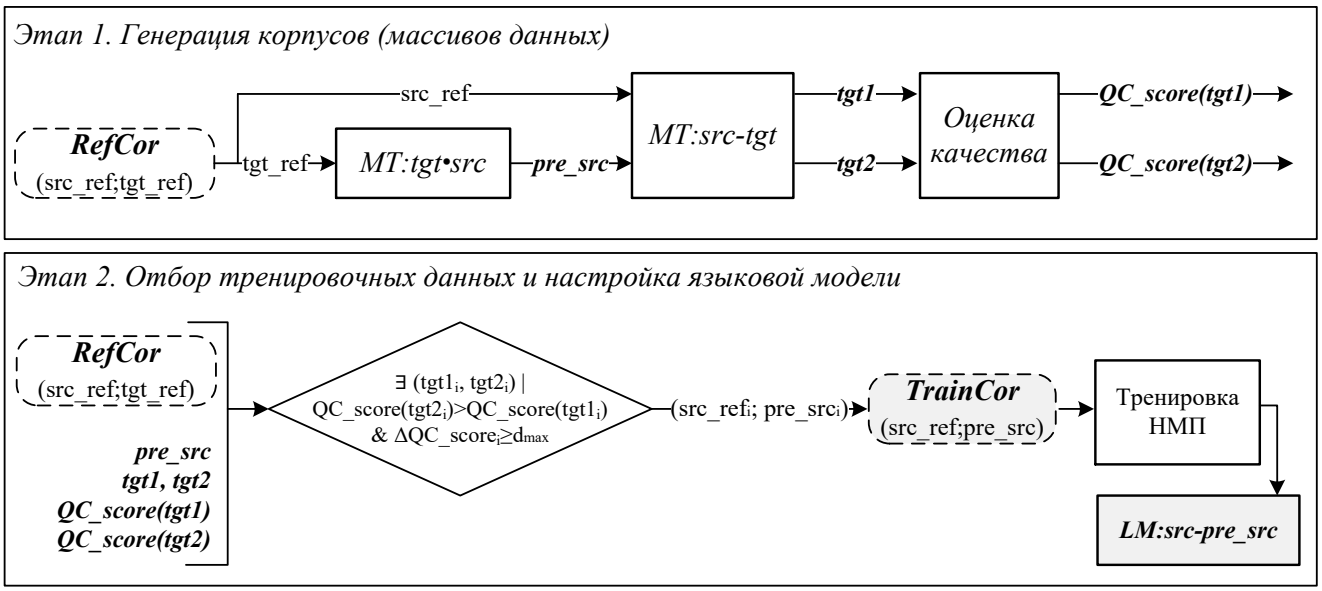

Рисунок 3.1 – Обучение модели оптимизационного предредактирования

Созданная модель может использоваться для оптимизационного предредактирования текстов как самостоятельный инструмент, так и в составе программного комплекса.

3.3 Обучение модели оценки сложности задачи перевода

Для обучения модели оценки сложности задачи перевода заданного текста txt_{iTXT} с языка $яз_{вх}$ на язык $яз_{вых}$ переводчиком $пер_{iПЕР}$ в соответствии с формализованными требованиями к переводу $TP|txt_{iTXT}$ требуются исходные данные в виде корпуса параллельных текстов следующей структуры:

TranslatorExpCor: [src; trg; ref], где *src* – это оригинал, т.е. текст на языке $яз_{вх}$, *trg* – это перевод (текст на языке $яз_{вых}$), выполненный переводчиком $пер_{iПЕР}$; *ref* – это контрольный перевод (текст на языке $яз_{вых}$), т.е. проверенный эталон.

Этап 1. Оценка качества перевода. На данном этапе для каждого фрагмента текста $i=\overline{1,N}$, где N – общее число записей в корпусе *TranslatorExpCor,* производится оценка $ОК_i$,при условии, что $ОК_i \in R$. Полученные значения записываются

отдельным столбцом *QC_score* в *TranslatorExpCor.* Столбцы *[trg; ref]* удаляются как избыточные.

Этап 2. Структурный анализ предложений исходного текста. На данном этапе необходимо получить вещественные значения свойств *СВ* текста в виде матрицы оценок C_{ouji} для каждого *i-го* фрагмента исходного текста. В зависимости от языка исходного текста состав свойств может отличаться, однако, в целом, в задачах обработки естественного языка свойства текста условно можно разделить на группы признаков: *общие* (количество символов/слов/строк и т.д., стиль, язык, домен приложения и пр.), *ОП ⊂ СВ; лексические* (процент покрытия текста лексическими минимумами, частотными списками, специфичность лексики и др.), *ЛП ⊂ СВ*; *морфологические* (лексические и грамматические свойства формы слова), *МП ⊂ СВ*; *синтаксические* (глубина глагольных и именных групп, связи между глаголами в предложениях), *СП ⊂ СВ*; *признаки, основанные на базовых подсчетах* (средняя длина слов и предложений и пр.), *БП ⊂ СВ*.

Для морфологического и синтаксического разбора текста используем схему Universal Dependencies [111], которая позволяет производить анализ отдельных слов в предложении и их взаимосвязей, применив преобразование, для вещественной оценки свойств всего текста.

Пусть $txt_{iTXT} = \{t_0, t_1, \ldots, t_m\}$, где t_m – это токен текста (слово или знак препинания), m – общее число токенов в заданном тексте; $UD = МП \cup СП = \{UD_0, UD_1, \ldots, UD_n\}$, где UD_n – это морфологическое или синтаксическое свойство токена согласно схеме Universal Dependencies, n – это общее число возможных морфологических и синтаксических свойств токена по схеме Universal Dependencies, тогда

$$\forall\ UD_n\ \exists\ C_{ouji,k} = \frac{\sum_l^m 1 | f(t_l)=UD_n}{m} \tag{76}$$

где $t_l \in txt_{iTXT}$, $f(t_l)$ – функция морфологического/синтаксического анализа.

Общие признаки, лексические признаки и признаки, основанные на базовых подсчетах, выбираются, формализуются и рассчитываются, исходя из особенностей языка $яз_{вх}$, и дополняют матрицу C_{ouji}. Далее результаты вещественной оценки свойств текста записываются в *TranslatorExpCor.* Столбец *[src]* удаляется как избыточный.

Этап 3. Регрессионный анализ. На данном этапе, данные *TranslatorExpCor* разбиваются на 2 выборки согласно принципу Парето*: TranslatorExpCor_train* (80%) для моделирования, *TranslatorExpCor_test* (20%) для валидации модели. Для выборки *TranslatorExpCor_train* проводится регрессионный анализ относительно целевой переменной *QC_score*.

Коэффициенты регрессионной модели составляют матрицу весов значимости оценок свойств текста $W_{оц}$.

Этап 4. Оценка сложности задачи перевода. Используя данные, полученные в результате регрессионного анализа и применив формулы (37), (38), (39) рассчитывается сложность задачи перевода *СлЗП* для каждого *i-го* фрагмента текста.

Затем происходит отбор фрагментов текста, которым соответствуют высокие значения $СлЗП_i$ по формуле (42) с учетом $СлЗП_{доп}$ – допустимого значения сложности задачи перевода.

Оптимизационное предредактирование текста производится в соответствии с имеющимися методами и алгоритмами оптимизационного предредактирования по критерию минимизации *СлЗП*.

Выводы по третьей главе

В рамках разработки методов реализации обучения моделей для оптимизационного предредактирования русскоязычных текстов впервые предложен алгоритм оценки сложности переводческой задачи для переводчика на основе его компетенции и специализации и параметров исходного текста, которая позволяет прогнозировать риски некачественного и/или несвоевременного решения задачи перевода; предложен новый алгоритм оценки русскоязычного текста по лексическим, синтаксическим и морфологическим признакам; предложена новая методика обучения модели для редактирования русскоязычных текстов, отличающаяся от существующих применением обратного перевода для сбора тренировочных данных и использованием критерия оптимизации для повышения качества машинного перевода на английский язык.

При реализации описанных методов обучения моделей следует учитывать следующие допущения:

1) Критерий качества перевода должен быть четко определен и формализован с возможностью получения вещественного нормированного значения. Могут применяться любые метрики оценки качества в зависимости от требований к качеству перевода.

2) Для тестирования МП необходим тренировочный корпус, включающий тексты на языке оригинала и перевод, принятый за эталон. В компаниях, внедривших ISO 17100 и CAT, процесс накопления тренировочных корпусов, включающих исходный текст, перевод, выполненный системой МП и проверенный перевод, утвержденный редактором, происходит автоматически в режиме реального времени.

ГЛАВА 4 ПРОГРАММНЫЙ КОМПЛЕКС ОПТИМИЗАЦИОННОГО ПРЕДРЕДАКТИРОВАНИЯ УЗКОСПЕЦИАЛЬНЫХ РУССКОЯЗЫЧНЫХ ТЕКСТОВ ДЛЯ ИХ ПЕРЕВОДА НА АНГЛИЙСКИЙ ЯЗЫК

4.1 Архитектура программного комплекса и его подсистем

Программный комплекс оптимизационного предредактирования, далее – программный комплекс, реализован на языке Python и состоит из трех основных подсистем: подсистемы тренировки языковой модели, подсистемы оценки сложности задачи перевода, подсистемы оптимизационного предредактирования русскоязычного текста и генерации машинного перевода на английский язык. Схематично архитектура программного комплекса представлена на рисунке 4.1.

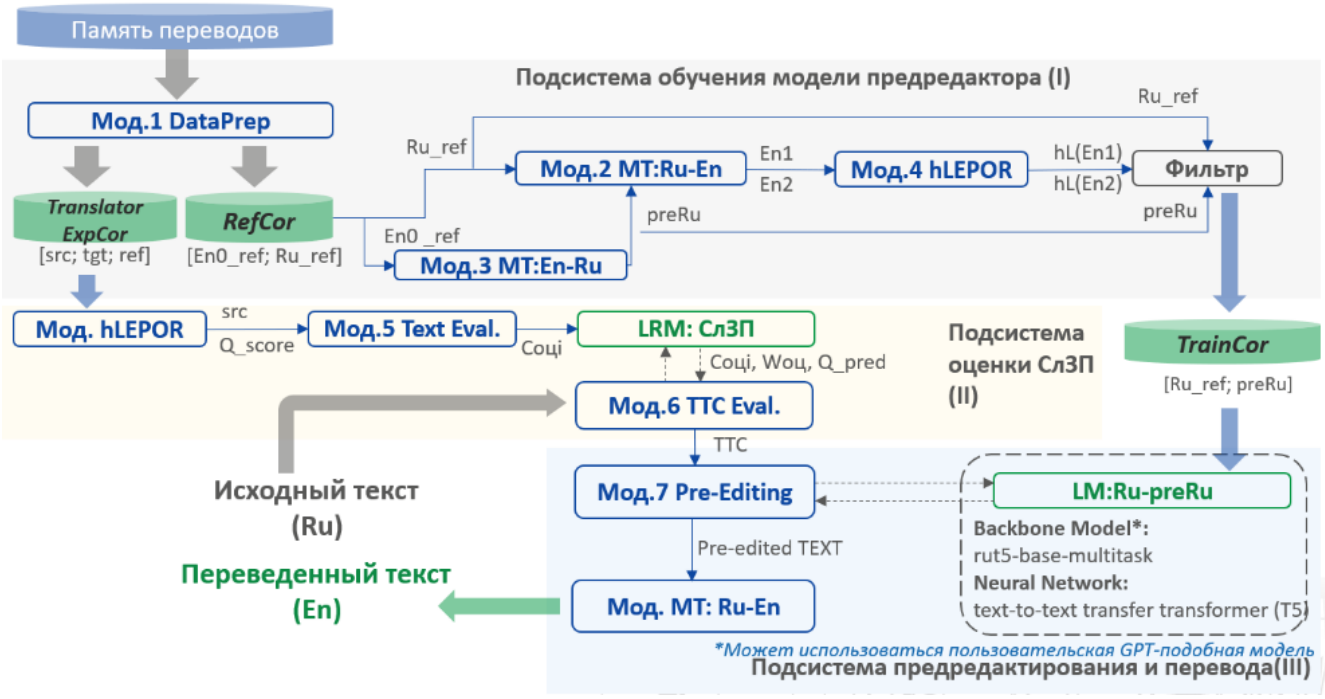

Рисунок 4.1 – Архитектура программного комплекса для повышения качества МП русскоязычных текстов на английский язык путем оптимизационного предредактирования

Подсистема тренировки языковой модели оптимизационного предредактирования русскоязычных текстов (I) состоит из пяти программных компонентов:

Мод.1 DataPrep – модуля подготовки «сырых данных», полученных из памятей переводов «Translation Memories» поставщика лингвистических услуг, который формирует корпуса TranslatorExpCor и RefCor для тренировки моделей оценки сложности задачи перевода и оптимизационного предредактирования соответственно; *Мод.2 MT:Ru-En*, *Мод.3 MT:En-Ru* – модулей МП (генератор перевода в языковой паре русский-английский и генератор в языковой паре английский-русский); *Мод.4 hLEPOR* – модуля оценки качества МП реализует алгоритм по метрике hLEPOR; модуля фильтрации данных, подходящих для тренировки модели, в котором производится отбор по условию повышения оценки качества после предредактирования. В результате обработки эталонного корпуса RefCor модулями системы 1 – 3 полученный тренировочный корпус TrainCor используется для обучения модели оптимизационного предредактирования *LM:Ru-preRu*.

Подсистема оценки сложности задачи перевода (II) состоит из трех модулей: модуля оценки качества перевода, выполненного системой МП, относительно эталонного по метрике hLEPOR; *Мод.5 Text Eval.* – препроцессора для взвешенной оценки свойств русскоязычного текста, включая морфологические, синтаксические, лексические и другие, всего 96 параметров; *Мод.6 TTC Eval.* – модуля оценки сложности задачи перевода с применением модели логистической регрессии *LRM:СлЗП*, которая определяет ожидаемое качество перевода полученного текста на основании взвешенной оценки его свойств и весов значимости оценок относительно системы МП.

Подсистема оптимизационного предредактирования и генерации МП (III) состоит из двух модулей: *Мод.7 Pre-editing* – модуля автоматического предредактирования русскоязычных текстов на основе модели *LM:Ru-preRu*, которая в качестве опорной использует модель русского языка rut5-base-multitask на основе нейронной сети типа *text-to-text transfer transformer* (T5), дообученной на корпусе параллельных русскоязычных текстов подзадачу перефразирования; модуля генерации МП с русского языка на английский язык на основе модели *Helsinki-NLP/opus-mt-ru-en*.

Программный комплекс реализован на языке Python, модули обучения моделей *LRM:СлЗП* и *LM:Ru-preRu* развернуты в Центре коллективного пользования научным

оборудованием «Центр обработки и хранения научных данных ДВО РАН» на базе ВЦ ДВО РАН – обособленном подразделении ХФИЦ ДВО РАН. Тип используемой ЭВМ: ПК с архитектурой x86, x86_64.

4.1.1 Модуль автоматической очистки сырых данных из памятей переводов CAT для тренировки языковой модели

Программный модуль *Мод.1 DataPrep* реализует алгоритмы автоматической обработки параллельных корпусов текстовых данных с целью их очистки и подготовки к тренировке моделей нейронного машинного перевода. Этапы очистки данных, реализуемые в программном модуле:

- удаление дубликатов;
- удаление тегов разметки текста;
- удаление строк, где оригинал совпадает с переводом;
- удаление слишком длинных и слишком коротких строк; удаление строк, в которых менее 30% алфавитных символов.

Программный модуль может использоваться как автономное приложение в задачах прикладной компьютерной лингвистики и повышения качества перевода.

4.1.2 Модули машинного перевода

В случаях, когда сниженные требования к переводу не оправдывают затраты на привлечение высококвалифицированных исполнителей, возможно использование систем МП. Но стоит учитывать, что доступные для общего пользователя системы, такие как Google Translator, DeepL, Яндекс Переводчик и прочие, имеют ограничения по объему бесплатного использования и их политика конфиденциальности, как правило, подразумевает сбор, хранение и обработку вводимой информации при бесплатном использовании, что не всегда допустимо для пользователя. Решить данную проблему возможно путем реализации программного парсера МП на основе доступных OpenSource моделей.

Для генерации МП с русского языка на английский язык и с английского языка на русский язык разработаны программные модули-парсеры [112] *Мод.2 MT:Ru-En*, *Мод.3 MT:En-R* с использованием модели нейронного МП на базе Transformers [113] и Open Source модели Helsinki-NLP от команды Hugging Face, участников

крупнейшего ежегодного соревнования разработчиком систем МП WMT [114], предварительно обученной на корпусе OPUS [115].

Парсеры получают в качестве исходных данных .xls-файл, в котором содержится построчно сегментированный текст. Используя настроенную модель МП, парсер анализирует текст и генерирует текст на требуемом языке. Результирующий массив строк, содержащий перевод, сохраняется в файл формата .xls. Разработка и инициализация парсеров МП производилась в среде Google Colab.

Разработанные парсеры могут использоваться как автономные риложения и позволяют генерировать до 10 000 страниц перевода с русского языка на английский язык в сутки на стороне сервера Google Colab. Использование OpenSource моделей и создание подобных парсеров МП на основе архитектуры Transformers может эффективно применяться с целью оптимизировать затраты и время на выполнение перевода в прикладных и исследовательских задачах различных областей знаний в условия сниженных требований к языковым и стилистическим нормам.

4.1.3 Модуль оценки качества машинного перевода

Модуль *Mod.4 hLEPOR* реализует алгоритм оценки качества МП. Для оценки МП выбрана метрика hLEPOR, которая является комбинацией существующих и доработанных факторов. Метрика показывает лучшие результаты оценки по сравнению с MPF, ROSE, METEOR, BLEU и TER и имеет наивысший балл корреляции Пирсона с человеческими суждениями по языковой паре английский-русский [116].

Оценка метрики производится путем сравнения сгенерированного МП (гипотезы) и эталонным переводом, выполненным человеком. Метрика изменяется в пределах от 0 до 1, где 0 – полное несовпадение гипотезы с эталоном, а 1 – полное совпадение.

4.1.4 Модуль препроцессинга текстовых данных для взвешенной оценки параметров русскоязычного текста

В общем смысле текст – это письменное сообщение, объективированное в виде письменного документа, состоящее из ряда высказываний, объединённых разными типами лексической, грамматической и логической связи, имеющее определённый моральный характер, прагматическую установку и соответственно

литературно обработанное. Текст обладает морфологическими, лексическими, синтаксическими признаками (объективные оценки), а также признаками, определяемыми в отношение субъекта учитывая его информационную (когнитивную) трудность [89]. Когнитивная трудность текста определяется на уровне реципиента его способностью идентифицировать семантические единицы в тексте.

Объективные оценки текста могут быть рассчитаны и использованы для статистического анализа и сравнения текстов, их кластеризации и классификации, а также при создании и исследовании цифровых двойников в системах обработки естественного языка.

Модуль препроцессинга текстовых данных для взвешенной оценки параметров русскоязычного текста [117] автоматически рассчитывает вещественную оценку признаков русскоязычного текста по 96 параметрам, разделенным на четыре группы:

1 *Общие признаки текста*, которые включают: длину предложения в символах; число токенов в предложении; долю слов в предложении; среднее количество слогов в слове; средняя длина слова в символах; долю токенов пунктуации; глубину дерева синтаксического разбора; индекс читабельности Флэша, адаптированной для русского языка И. В. Оборневой [88]. Индекс читабельности Флэша основан на гипотезе, что чем меньше слов в предложениях и чем короче эти слова, тем проще текст. Данный индекс принимает значение от 1 до 100, где тексты с индексом меньше 30 очень трудно читать, с индексом 70 и выше должны быть легко читаемыми [118]. Расчет индекса читабельности Флэша производится по формуле:

$$R_f = 206{,}835 - (1{,}3 * mW) - (60{,}1 * mS), \quad (77)$$

где mS – среднее количество слогов в словах, mW – количество слов в предложении.

2 *Морфологические и синтаксические признаки текста*, для оценки которых используется разбор по схеме универсальных зависимостей [104]. Морфологическая спецификация слова в схеме универсальных зависимостей состоит из трех частей: лемма слова, тег части речи и морфологические признаки, которые определяют лексические и грамматические свойства формы слова. В данном случае оценивается доля слов исследуемой части текста по каждому тегу части речи, как наиболее информативный морфологический признак. Универсальные зависимости имеют

фиксированный список тегов, включающий 17 частей речи. При синтаксическом разборе текста, каждая его единица (токен) маркируется соответствующим тегом отношения «rel=». Фиксированный список синтаксических связей включает 64 тега. Морфологический и синтаксический разбор выполняется следующим образом:

1) Разбиение текста на токены.

2) Определение для каждого токена значений свойств pos (часть речи) и rel (роль в предложении).

3) Подсчет количества частей речи и числа токенов по каждому тегу в каждой строке.

4) Расчет для всех тегов отношения количества токенов с соответствующим тегом к общему числу токенов в предложении.

3 *Лексические признаки текста:* доля именованных сущностей; доля слов, входящих в частотные списки наиболее употребимой лексики русского языка (для ТОП-1000, ТОП-3000, ТОП-5000 и ТОП-10000); доля слов с неизвестным значением частотности, то есть тех, которых нет в частотных списках; доля слов с коэффициентом вариации *Жуйана* < D_{max}, значение которого зависит от того, является ли слово термином (коэффициент Жуйана является лучшим из известных в настоящее время способов измерить, насколько общеупотребительным является слово, или, напротив, насколько оно специфично для отдельных предметных областей [119]). Для оценки лексических признаков используется Национальный частотный словарь русской лексики [110], который был предварительно обработан и отсортирован по значению частотности слов при помощи средств языка Python в среде программирования Google Colab. Ориентируясь на распределение вариации Жуайна для Национального частотного словаря русской лексики, принят коэффициент D_{max}=79. Оценка лексических признаков производится по токенам, включающим слова, которые предварительно были лемматизированы.

Для автоматической оценки русскоязычного текста по описанным группам параметров был разработан программный модуль на языке Python. Исходными данными для анализа являются файлы .xls, в которых собраны тексты, сегментированные по предложениям. Морфологический и синтаксический анализ выполняется при

помощи библиотеки natasha [120] для моделирования систем обработки естественного языка на основе глубокого обучения для русского языка.

Разработанный программный модуль *Мод.5 Text Eval.* является универсальным и может быть использован в качестве автономного приложения для анализа русскоязычных текстов любой тематики. Тестирование программного модуля показало, что он может использоваться для анализа корпусов текстов большого объема. Результаты оценки текстов могут быть использованы в задачах статистического анализа, сравнения текстов, выявления общих признаков, кластеризации и других задачах компьютерной лингвистики.

4.1.5 Модуль вероятностной оценки сложности переводческой задачи для систем машинного перевода

Модуль вероятностной оценки сложности переводческой задачи для систем машинного перевода *Мод.6 TTC Eval.* позволяет рассчитывать оценку сложности задачи его перевода на английский язык с использованием модели МП [121].

Модуль обращается к модели логистической регрессии *LRM:СлЗП*, которая обучается на массиве данных, включающих исходный текст, эталонный перевод и перевод, выполненный системой машинного перевода. Обученная модель решает задачу классификации, рассчитывая вероятность попадания переведенного текста в класс «качественный перевод» в соответствии с требованиями пользователя.

Модель обучается на данных полученных в результате предварительного обучения на массиве исходных данных, обработанных в модуле препроцессинга текстовых данных для взвешенной оценки параметров русскоязычного текста.

Перед обучением модели логистической регрессии *LRM:СлЗП* был проведен анализ полученных оценок параметров текста (факторов) на нормальность распределения показала наличие большого количества факторов с экспоненциальным распределением значений и факторов, большая часть наблюдений по которым равна 0 (пример представлен на рисунке 4.2). Будем учитывать это при оценке точности модели, так как нормализация и шкалирование данных не позволит устранить нулевые значения, а балансировка данных по нулевым значениям может привести к значительным потерям данных по другим факторам.

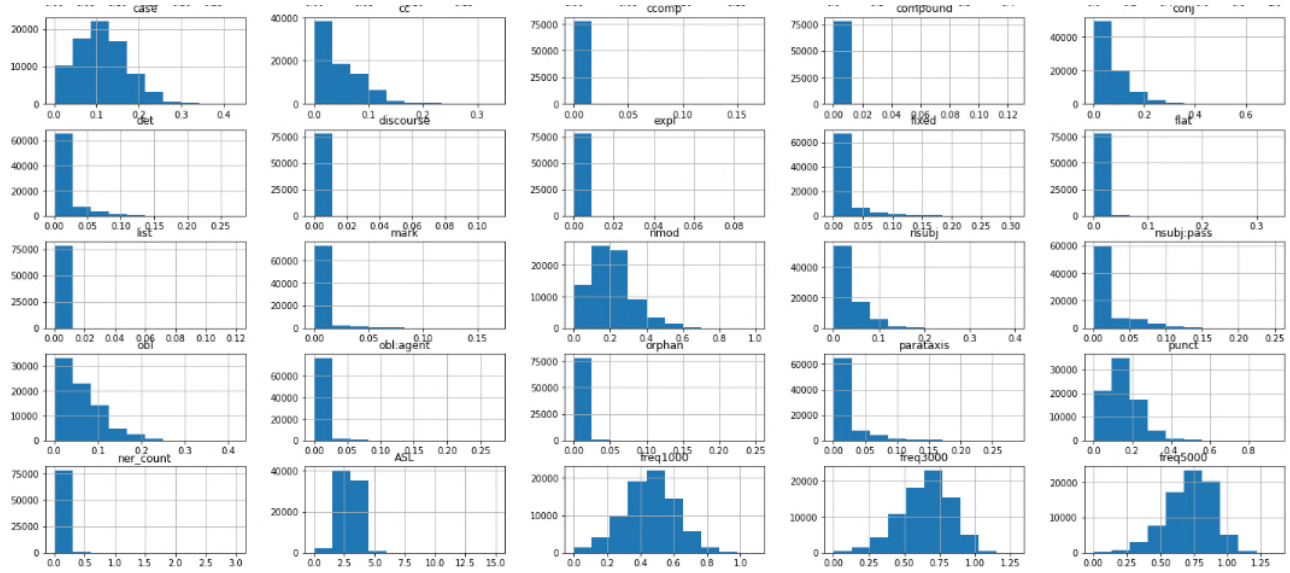

Рисунок 4.2 – Гистограммы распределения данных для части факторов

Для поиска весов значимости параметров текста и получения многофакторного уравнения поиска теоретического качества перевода использовалась модель, реализованная в библиотеках statsmodels.api и sklearn для языка Python. Результаты моделирования представлены на рисунке 4.3.

	coef	std err	z	P>|z|	[0.025	0.975]
const	-2.1237	0.119	-17.917	0.000	-2.356	-1.891
ADJ	0.8890	0.145	6.113	0.000	0.604	1.174
ADP	-1.1872	0.191	-6.222	0.000	-1.561	-0.813
ADV	-1.1501	0.323	-3.562	0.000	-1.783	-0.517
DET	1.1331	0.371	3.055	0.002	0.406	1.860
NUM	4.3565	0.227	19.228	0.000	3.912	4.801
PRON	-3.2955	0.419	-7.869	0.000	-4.116	-2.475
PROPN	-0.9230	0.136	-6.799	0.000	-1.189	-0.657
VERB	-0.9131	0.234	-3.898	0.000	-1.372	-0.454
X	2.3469	0.409	5.733	0.000	1.545	3.149
acl	-1.3931	0.356	-3.916	0.000	-2.090	-0.696
aux	5.1366	1.905	2.696	0.007	1.403	8.870
conj	2.2937	0.178	12.850	0.000	1.944	2.644
csubj	-4.6383	0.760	-6.099	0.000	-6.129	-3.148
flat	4.3092	1.322	3.261	0.001	1.719	6.899
flat:name	-4.4942	1.231	-3.650	0.000	-6.908	-2.081
list	13.0258	4.971	2.620	0.009	3.283	22.768
nmod	-1.6506	0.131	-12.597	0.000	-1.907	-1.394
nsubj:pass	-0.7066	0.310	-2.276	0.023	-1.315	-0.098
obj	-1.3491	0.261	-5.175	0.000	-1.860	-0.838
obl	-0.7080	0.224	-3.161	0.002	-1.147	-0.269
obl:agent	3.2594	0.843	3.865	0.000	1.606	4.912
parataxis	3.7789	0.360	10.506	0.000	3.074	4.484
punct	0.3628	0.168	2.166	0.030	0.034	0.691
xcomp	1.3277	0.380	3.490	0.000	0.582	2.073
freq10000	1.6783	0.064	26.415	0.000	1.554	1.803
tokens_count_log	0.3153	0.018	17.294	0.000	0.280	0.351

Рисунок 4.3 – Коэффициенты логистической регрессии

В итоговую модель вошли только те факторы (параметры исходного текста), для которых P-значения показывают высокую значимость. Таким образом, были получены коэффициенты уравнения поиска теоретического качества перевода, выполненного тестируемым переводчиком [122]. Всего выявлено 26 параметров, из которых 12 имеют отрицательную зависимость с потенциальной оценкой качества перевода и являются мешающими (таблица 4.1).

Таблица 4.1 – Список значимых факторов модели логистической регрессии

Фактор	**Коэф-т**	**Описание фактора (значение тега)**
list	13,0258	Используется для цепочек сопоставимых элементов, причем в списках с более чем двумя элементами все элементы списка должны модифицировать первый элемент.
aux	5,1366	Функциональное слово, связанное с глагольным предикатом и выражающее такие категории, как время, настроение, аспект, голос или эвиденциальность.
csubj	-4,6383	Синтаксический субъект предложения, т.е. субъект сам является предложением.
flat:name	-4,4942	Уточнение тега flat, используемое для имен.
NUM	4,3565	Слово, функционирующее чаще всего как определитель, прилагательное или местоимение, которое выражает число и отношение к числу, такое как количество, последовательность, частота или дробь.
flat	4,3092	Одно из трех отношений для многословных выражений (MWEs) в схема универсальных зависимостей (два других - фиксированные и составные), используется для (безглавых) полуфиксированных MWE, таких как имена и даты.
parataxis	3,7789	Связь между словом (часто главным предикатом предложения) и другими элементами, например, сентенциальной скобкой или клаузой после ":" или ";", расположенными рядом без явного согласования, подчинения или аргументной связи с главным словом.
obl:agent	3,3594	Используется для обозначения агентов пассивных глаголов.
PRON	-3,2955	Слова, заменяющие существительные или фразы существительных, значение которых можно восстановить из лингвистического или экстралингвистического контекста.
X	2,3469	Слова, которым по каким-то причинам не может быть присвоена категория части речи.
conj	2,2937	Отношение между двумя элементами, связанными координационным союзом, таким как «и», «или» и т.д.
freq10000	1,6783	Вхождение слова в ТОП-10000 наиболее частотных слов русского языка.
nmod	-1,6506	Используется для номинативных зависимостей от другого существительного или именной фразы.
acl	-1,3931	Обозначает конечные и не конечные предложения, которые модифицируют номинальное.

Фактор	Коэф-т	Описание фактора (значение тега)
obj	-1,3491	Второй по значимости аргумент глагола после субъекта.
xcomp	1,3277	Предикативное или клаузальное дополнение без собственного субъекта.
ADP	-1,1872	Общий термин для предлогов и постпозиций.
ADV	-1,1501	Наречия, т.е. слова, которые обычно изменяют глаголы по таким категориям, как время, место, направление или тональность. Они также могут модифицировать прилагательные и другие наречия.
DET	1,1331	Слова, которые изменяют существительные или фразы существительных и выражают референцию фразы существительного в контексте.
PROPN	-0,9230	Имя собственное – это существительное (или слово номинативного содержания), которое является именем (или частью имени) конкретного лица, места или предмета.
VERB	-0,9131	Глагол, является членом синтаксического класса слов, которые обычно обозначают события и действия, могут составлять минимальный предикат в клаузе и регулируют количество и типы других составляющих, которые могут встречаться в предложении.
ADJ	0,8890	Прилагательные, слова, которые обычно изменяют существительные и указывают на их свойства или признаки.
obl	-0,7080	Используется для основной части речи (существительное, местоимение, фраза существительного), функционирующего как неосновной (косвенный) аргумент или дополнение.
nsubj:pass	-0,7066	Пассивный номинальный субъект – это фраза существительного, которая является синтаксическим субъектом пассивного предложения.
punct	0,3628	Используется для любого знака препинания в предложении, если пунктуация сохраняется в типизированных зависимостях.
tokens_count_log	0,3153	Натуральный логарифм количество токенов в предложении.

Данные в таблице отсортированы по убыванию абсолютной величины коэффициента логистической регрессии. По таблице мы можем оценить какие признаки текста имеют больший вес при решении задачи бинарной классификации, а также характер зависимости (положительная/ отрицательная).

Выполняя предредактирование исходного текста с целью оптимизации данных параметров возможно повысить качество МП. Кроме того, 14 выявленных параметров имеют положительную зависимость с потенциальной оценкой качества перевода, что тоже может использоваться в оптимизационном предредактировании

для снижения влияния мешающих параметров. Проверка качества полученной модели проводилась на основе показателя ROC-AUC [123].

ROC-AUC = 0,6456 показывает, что на основании признаков исходного текста без оценки его семантики, возможно предсказывать ожидаемое качество МП. Максимальная вероятность полученных предсказаний по тестовой выборке равна 0,9765, что говорит нам о потенциале доработки данного классификатора и повышения его точности, с учетом разреженности пространства факторов и распределения их значений. Принимая во внимание допущение, что автоматическая оценка качества перевода имеет некоторую погрешность, можно говорить о состоятельности разработанной модели и предложенного алгоритма оценки сложности переводческой задачи перевода текстов нефтегазовой тематики для выбранной системы МП.

4.1.6 Предредактор русскоязычных узкоспециальных текстов для систем машинного перевода

Модуль автоматического предредактирования русскоязычных текстов *Мод.7 Pre-editing* получает в качестве входных данных сегменты текста, у которых оценка сложности задачи перевода выше допустимой. Модуль работает только совместно с тренировочным модулем языковой модели для перефразирования русскоязычных технических текстов *LM:Ru-preRu*, обращаясь к модели оптимизационного предредактирования для генерация оптимизированных текстов.

Программный модуль *LM:Ru-preRu* реализует алгоритм инициализации и тренировки языковой модели на основе архитектуры transformers для перефразирования русскоязычных технических текстов с сохранением семантической идентичности для использования в задачах повышения качества машинного перевода.

В рамках настоящего исследования используется опорная модель расширенной архитектуры T5-base, предобученная для решения различных задач генерации текстовых последовательностей, со следующими характеристиками: количество слоев нейросети трансформера – 16 (по 8 слоев энкодера и декодера); размер матриц весов нейронов – 1024x1024; общее число нейронов – 125 441 600; исходных объем данных предварительного обучения – более 1,56Т слов.

Исходными данными для дообучения модели задаче оптимизационного предредактирования является *TrainCor* – корпус параллельных двуязычных текстов нефтегазовой тематики для тренировки языковых моделей в задачах перефразирования узкоспециальных технических русскоязычных текстов и повышения качества их перевода на английский язык, полученный в результате очистки и подготовки данных из памятей переводов средств автоматизированной поддержки переводческой деятельности.

Программный модуль *Mod.7 Pre-editing* в комплексе с *LM:Ru-preRu* может быть использован как автономное приложение в задачах оптимизационного предредактирования русскоязычных текстов либо дообучен при помощи соответствующих наборов данных для решения задач перефразирования и/или симплификации русскоязычных текстов в соответствии с заданными критериями.

4.2 Данные для обучения и тестирования программного комплекса

Исходные данные для обучения моделей и тестирования в виде русскоязычных узкоспециальных технических текстов, переведенных на английский язык, предоставлены ведущей компанией по оказанию лингвистических услуг в области технического перевода в Хабаровском крае ООО «Агентство переводов «ФИАС-Амур» в объеме ~60 000 ст. стр. текста (1 ст. стр. = 1800 знаков с пробелами). Из предоставленных данных сформированы корпуса *RefCor* объемом 139 438 семплов и *TranslatorExpCor* объемом 83 543 семплов. После обработки корпуса *RefCor* в подсистеме тренировки языковой модели оптимизационного предредактирования русскоязычных текстов в корпус *TrainCor* вошло 88 631 семпл. Тестовая выборка для оценки работы системы *TestCor* составила 16 707 семплов.

4.2.1 База данных структурного анализа предложений технических русскоязычных текстов

Путем обработки корпуса *TranslatorExpCor* модулями оценки качества машинного перевода *Mod. hLEPOR* и препроцессинга текстовых данных для взвешенной оценки параметров русскоязычного текста *Mod.6 TTC Eval.* получена база

данных структурного анализа предложений технических русско-язычных текстов, которая содержит 83543 русскоязычных предложений технического текста объемом от 50 до 300 символов, для которых выполнен расчет вещественных параметров по морфологическим, синтаксическим, лексическим и прочим признакам. Всего база данных содержит расчёты для 96 признаков. Морфологические и синтаксические признаки рассчитаны на основе универсальных зависимостей с учетом доли токенов определенного признака в предложении.

Формат: файл .xlsx. *Название*: DB_RUS_Tech_texts_UD. *Объем*: 25,4 МБ.

База данных содержит русскоязычные предложения технического текста объемом от 50 до 300 символов (Ru_test). Объем русскоязычного текста составляет ок. 5900 ст. стр. (1 ст.стр. = 1800 знаков с пробелами). Поля базы данных:

- En0_test – эталонный ручной перевод, проверенный редактором;
- En1 – машинный перевод;
- h_Lepor – оценка машинного перевода, характеризующая близость машинного перевода эталонному, где 0 – минимальная близость; 1 – максимальная;
- h_Lepor_class – переменная, классифицирующая предложения по признаку оценки машинного перевода на 4 класса: 1 для hLEPOR < 0,25; 2 – 0,25 < hLEPOR < 0,5; 3 – 0, 5 < hLEPOR < 0,75; 4 –hLEPOR > 0,75;
- длина предложения в символах (len_ru);
- число токенов в предложении (tokens_count);
- доля слов в предложении (ASW_per_tokens);
- среднее количество слогов в слове (ASL);
- средняя длина слова в символах (ASW_mean_len);
- глубина дерева синтаксического разбора (uniqe_had_id_count);
- индекс читабельности Флэша (flesh_readability).
- Морфологические признаки текста: ADJ, ADP, ADV, AUX, CCONJ, DET, INTJ, NOUN, NUM, PART, PRON, PROPN, PUNCT, SCONJ, SYM, VERB, X (всего 17).
- Синтаксические признаки текста: acl, acl:relcl, advcl, advmod, advmod:emph, advmod:lmod, amod, appos, aux, aux:pass, case, cc, cc:preconj, ccomp, clf, compound, compound:lvc, compound:prt, compound:redup, compound:svc, conj,

cop, csubj, csubj:outer, csubj:pass, dep, det, det:numgov, det:nummod, det:poss, discourse, dislocated, expl, expl:impers, expl:pass, expl:pv, fixed, flat, flat:foreign, flat:name, goeswith, iobj, list, mark, nmod, nmod:poss, nmod:tmod, nsubj, nsubj:outer, nsubj:pass, nummod, nummod:gov, obj, obl, obl:agent, obl:arg, obl:lmod, obl:tmod, orphan, parataxis, punct, reparandum, vocative, xcomp (всего 64).

- доля именованных сущностей (ner_count);
- доля слов, входящих в частотные списки наиболее употребимой лексики русского языка (freq1000 для ТОП-1000, freq3000 для ТОП-3000, freq5000 для ТОП-5000 и freq10000 для ТОП-10000);
- доля слов с неизвестным значением частотности, то есть тех, которых нет в частотных списках (no_freq);
- доля слов с коэффициентом вариации Жуйана менее 79 (D_trg).

База данных может использоваться для тренировки моделей классификации, кластеризации и регрессии в решении задачи прикладной компьютерной лингвистики.

4.2.2 Корпус параллельных двуязычных текстов нефтегазовой тематики для тренировки языковых моделей в задачах перефразирования узкоспециальных технических русскоязычных текстов и повышения качества их перевода на английский язык

Корпус *TrainCor* – это корпус параллельных двуязычных текстов нефтегазовой тематики для тренировки языковых моделей в задачах перефразирования узкоспециальных технических русскоязычных текстов и повышения качества их перевода на английский язык, который включает в себя параллельные тексты нефтегазовой тематики, сегментированные на отдельные предложения, и имеет следующую структуру: исходный текст на русском языке, отредактированный вариант исходного текста на русском языке, эталонный перевод на английский язык, проверенный редактором; машинный перевод исходного русскоязычного текста; машинный перевод отредактированного русскоязычного текста. Кроме того, корпус включает показатели качества машинного перевода, рассчитанные по методике hLEPOR и оценивающие близость машинного перевода по отношению к эталонному в диапазоне от 0 до 1. Всего корпус содержит 88631 уникальных записей.

Формат: файл .xlsx. *Название*: DB_TrainCor_RU-preRU. *Объем*: 26,9 МБ

Область применения корпуса – тренировка языковых моделей в задачах обработки русскоязычных текстов и повышения качества машинного перевода на английский язык.

4.3 Тестирование программного комплекса

4.3.1 Постановка задачи тестирования

Целью тестирования является проверка работоспособности программного комплекса для повышения качества машинного перевода узкоспециальных технических текстов путем автоматического оптимизационного предредактирования.

Объект тестирования: программный комплекс оптимизационного предредактирования русскоязычных текстов и их перевода на английский язык (далее – программный комплекс) [124].

Предмет тестирования: качество машинного перевода на английский язык, полученного с помощью программного комплекса.

Гипотеза тестирования: использование оценки сложности задачи перевода в качестве критерия для отбора частей текста для оптимизационного предредактирования и их последующее предредактирование позволяют повысить качество машинного перевода.

4.3.2 Начальные условия и границы проведения тестирования

Для проведения тестирования используется случайная выборка реальных данных из предварительно подготовленных и очищенных памятей переводов поставщика лингвистических услуг. *Объем тестовой выборки*: 16707 семплов. *Сложность данных*: предложения длиной от 5 до 61 токенов.

Предметная область: узкоспециальные технические русскоязычные тексты по тематикам: нефтегазодобыча и переработка, морские платформы, транспортировка нефти и газа, техническая документация проектов строительства нефтегазовых терминалов.

В рамках тестирования примем, что минимально допустимая сложность задачи перевода $СлЗП_{доп} = 1{,}43$. Так как, согласно уравнению (39), сложность задачи

перевода обратно пропорциональна вероятности получения перевода требуемого качества, при $СлЗП_{доп}$ = 1,43 данная вероятность будет составлять 0,7.

Качество перевода оценивается с использованием алгоритма hLEPOR, рассчитывающего близость сгенерированной последовательности токенов эталону в диапазоне от 0 до 1, где 0 – полное несовпадение; 1 – полное совпадение. Эталоном выступает ручной перевод, проверенный квалифицированным редактором переводов.

Характеристики вычислительной системы: процессор – Intel(R) Core(TM) i7-4770 CPU, 3.40GHz; оперативная память – 32,0 ГБ; тип операционной системы – 64-разрядная ОС.

4.3.3 Методология и план тестирования

Тестирование включает три этапа:

1 Машинный перевод тестовой выборки и оценка его качества алгоритмом hLEPOR → получение оценки *hLEPOR(En1)*.

2 Оптимизационное предредактирование тестовой выборки → машинный перевод предредактированных текстов → получение оценки *hLEPOR(En2)*.

3 Оценка сложности задачи перевода тестовой выборки *СлЗП(Ru)* → применение оптимизационного предредактирования только к тем семплам, для которых $СлЗП > СлЗП_{доп}$ → оценка СлЗП после оптимизационного предредактирования *СлЗП(preRu)* машинный перевод → получение оценки *hLEPOR (TTC/PE)*.

4 Сравнение и анализ полученных результатов.

Схематичное представление методики тестирования языковой модели и модуля оптимизационного редактирования представлено на рисунке 4.4.

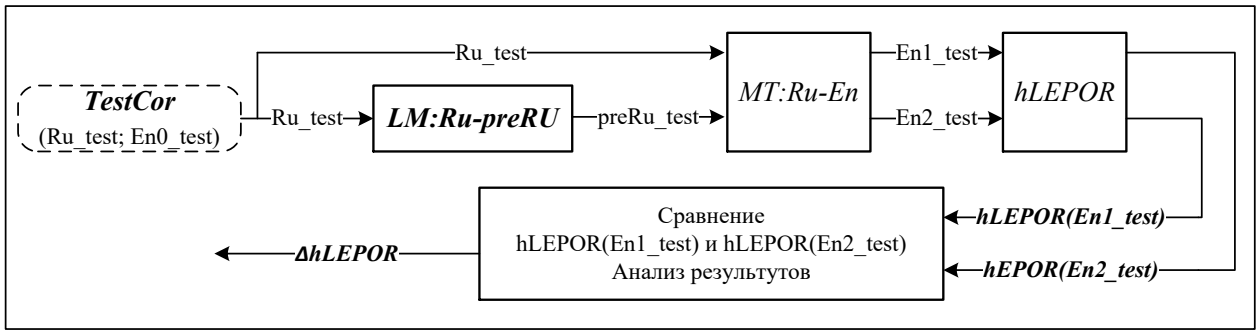

Рисунок 4.4 – Схематичное представление методики тестирования языковой модели и модуля оптимизационного редактирования

4.3.4 Результаты тестирования

Результаты, полученные в ходе тестирования, представлены на рисунках 4.5-4.7 и в таблице 4.2.

Принятые обозначения:

- mean – математическое ожидание;
- std – среднеквадратичное отклонение;
- min – минимальное значение выборки;
- 25% – значение, меньше которого 25% значений выборки;
- 50% – медиана, т.е. значение, меньше и больше которого 50% значений выборки;
- 75% – значение, меньше которого 75% значений выборки;
- max – максимальное значение в выборке.

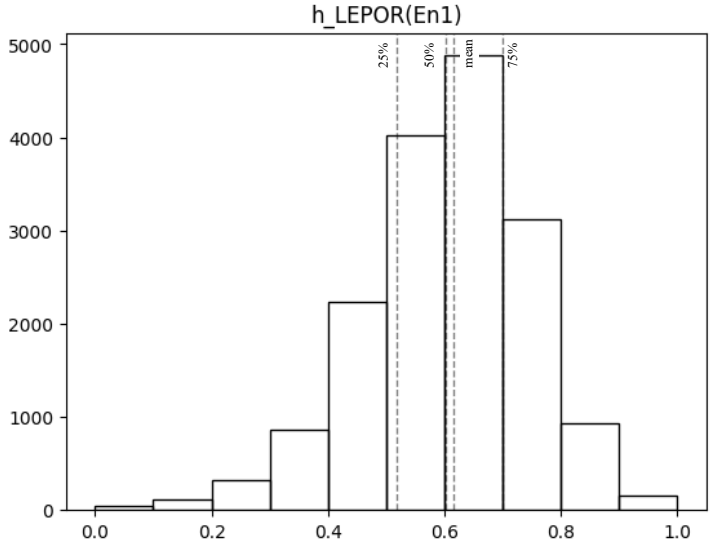

Рисунок 4.5 – Оценка качества машинного перевода до применения оптимизационного предредактирования

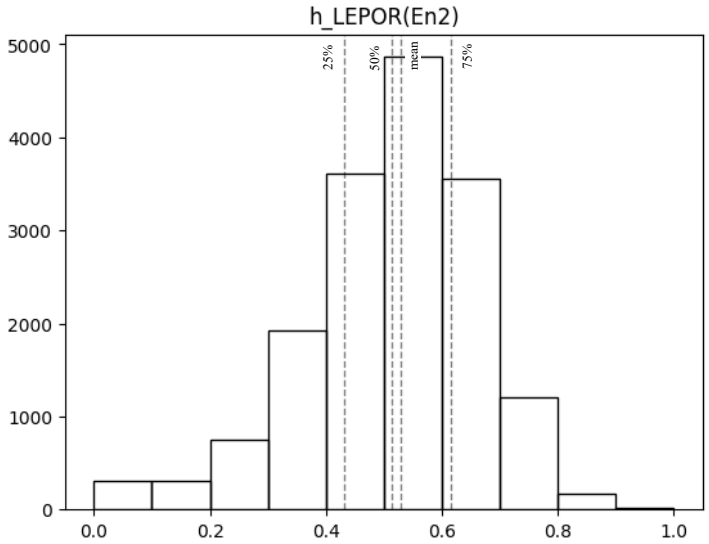

Рисунок 4.6 – Оценка качества машинного перевода после применения оптимизационного предредактирования ко всем семплам тестовой выборки

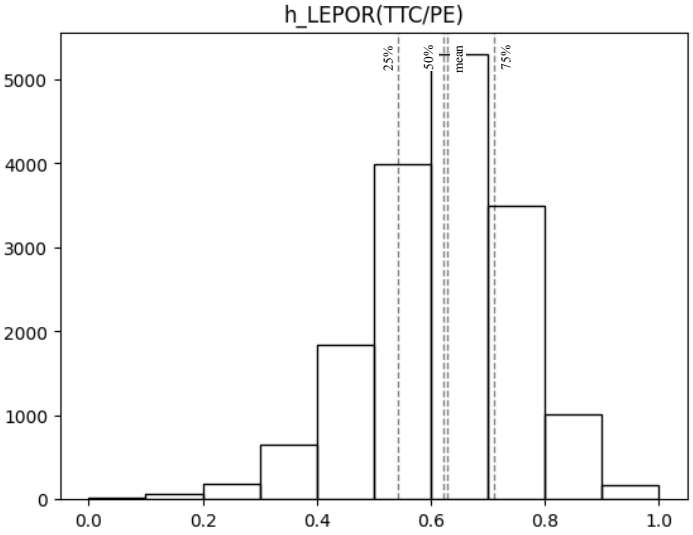

Рисунок 4.7 – Оценка качества машинного перевода после применения оптимизационного предредактирования к семплам тестовой выборки, отобранным по критерию сложности задачи перевода

Таблица 4.2 – Результаты тестирования программного комплекса

	h_LEPOR (En1)	**h_LEPOR (En2)**	**h_LEPOR (TTC/PE)**	**СлЗП (Ru)**	**СлЗП (preRu)**
mean	0,603154	0,51325	0,621428	2,058798	1,082323
std	0,142844	0,151931	0,13113	0,721968	0,053624
min	0	0	0	1,009145	1
25%	0,518918	0,430738	0,541798	1,650413	1,040528
50%	0,614855	0,529115	0,629695	1,885655	1,074627
75%	0,70099	0,61654	0,710368	2,23849	1,115369
max	1	1	1	20,19173	1,325531

Результаты тестирования показывают, что применение оптимизационного предредактирования без оценки его необходимости ведет к снижению общего качества выполненного перевода в среднем на 15%. Целесообразно использовать критерии отбора сегментов для оптимизационного предредактирования.

Использование оценки сложности задачи перевода в качестве критерия отбора сегментов и их последующего оптимизационного предредактирования позволяет повысить качество перевода. В рассматриваемой выборке общее среднее повышение качества перевода составило 3% при снижении сложности задачи перевода в среднем на 48%.

Рассмотрим результаты применения оптимизационного предредактирования. С использованием оценки сложности задачи перевода было отобрано 4071 семпл для оптимизационного предредактирования (24,4% тестовой выборки).

Примеры оптимизационного предредактирования и его влияния на сложность задачи перевода для системы МП представлены в таблице 4.3. Результаты перевода текстов после оптимизационного предредактирования на английский язык представлены в таблице 4.4.

Таблица 4.3 – Оптимизационное предредактирование русскоязычных текстов

Семпл	*Исходный текст (Ru_ref)*	*Текст после предредактирования (preRU)*	*ΔСлЗП*
1	Высоковольтные испытания проводятся по отдельно разрабатываемой и утверждаемой «Программе проведения высоковольтных испытаний кабеля 110 кВ».	Испытания на высоковольтные кабели проводятся в соответствии с отдельно разработанной и утвержденной программой испытаний высоковольтных кабелей 110 кВ.	-1,356
2	Оборудование должно быть рассчитано на двойные фидеры, а	Оборудование должно быть способно управлять двумя фидерами, в	-0,444

Семпл	*Исходный текст (Ru_ref)*	*Текст после предредактирования (preRU)*	*ΔСлЗП*
	если такое оборудование отсутствует, в центральном шкафу предусматривают установку контроллера автоматического ввода резерва.	случае отсутствия такого оборудования в центральном шкафу должен быть установлен переключатель ввода резерва.	
3	ТУ на поставку включают в себя, помимо прочего, следующее:	Спецификация покупки должна содержать и не ограничиваться:	-0,167
4	По результатам месяца подготовка отчета (10 число) по отклонениям от намеченного графика.	Отчет о ходе месяца (10 числа) о отклонениях от плана работы.	-0,271

Таблица 4.4 – Результаты машинного перевода на английский язык

Семпл	*МП исходного текста (En1)*	*МП после предредактирования (En2)*	*ΔhLEPOR*
1	The high voltage tests are conducted on a separate design and approval of the 110 kV high voltage test programme.	High voltage cables shall be tested according to a separately developed and approved 110 kV high voltage cables programme.	0,271
2	The equipment shall be designed for double feeders, and if such equipment is not available, an automatic backup controller shall be installed in the central cabinet.	The equipment shall be capable of controlling two feeders, in the absence of such equipment, a standby switch shall be installed in the central cabinet.	0,146
3	TA for supply includes, inter alia, the following:	The purchase specification shall contain and not be limited to:	0,135
4	Based on the month's results, the report (10 times) is based on deviations from the schedule.	Monthly progress report (10th) on deviations from the workplan.	0,131

Результаты оценки качества после оптимизационного предредактирования представлены в таблице 4.5.

Таблица 4.5 – Результаты применения оптимизационного предредактирования

	h_LEPOR (En1)	**h_LEPOR (En2)**	**ΔhLEPOR**	**СлЗП (Ru)**	**СлЗП (preRu)**
mean	0,508044336	0,583044424	0,075	2,111294	1,075499
std	0,140363512	0,128616684	0,0739	0,75958	0,052124
min	0	0,10348	1E-05	1,431731	1
25%	0,4253	0,506915	0,02056	1,692549	1,035022
50%	0,52027	0,59196	0,05472	1,928669	1,067745
75%	0,60549	0,67087	0,104235	2,314225	1,107176
max	0,99834	1	0,58203	20,19173	1,310526

Показано, что качество перевода отдельных сегментов, подвергшихся оптимизационному предредактированию, в среднем, возросло на 15%. Максимальное повышение качества перевода составило 30% в отдельных сегментах.

Результаты тестирования программного комплекса подтверждают работоспособность программного комплекса в части повышения качества машинного перевода узкоспециальных технических текстов путем автоматического оптимизационного предредактирования и использования оценки сложности задачи перевода в качестве в качестве критерия оптимизации. Это открывает широкие возможности для практического применения разработанной оценки при принятии решений в переводческих проектах, в том числе при использовании систем МП, а также использования оценки сложности задачи перевода в качестве критерия оптимизации. При этом программный комплекс в совокупности описанных подсистем и программных модулей имеет потенциал к доработке и повышении эффективности.

4.4 Внедрение программного комплекса в контур автоматизации процессов переводческой деятельности

Программный комплекс был внедрен в практическую деятельность ООО «Агентство переводов «ФИАС-Амур» (г. Комсомольск-на-Амуре), далее – Агентство переводов.

Целью внедрения программного комплекса в деятельность Агентства переводов является повышение качества машинного перевода, снижение затрат на его производство за счет повышения производительности редакторов переводов.

Внедрение программного комплекса проходило в несколько этапов:

1) Анализ бизнес-процесса осуществления перевода «как есть» с учетом средств автоматизации и поддержки процесса.

2) Внедрение программного комплекса в контур автоматизации бизнес-процесса перевода.

3) Обучение сотрудников.

4) Оценка и анализ результатов внедрения.

Схематично, процесс осуществления перевода в Агентстве переводов до внедрения программного комплекса представлен на рисунке 4.8.

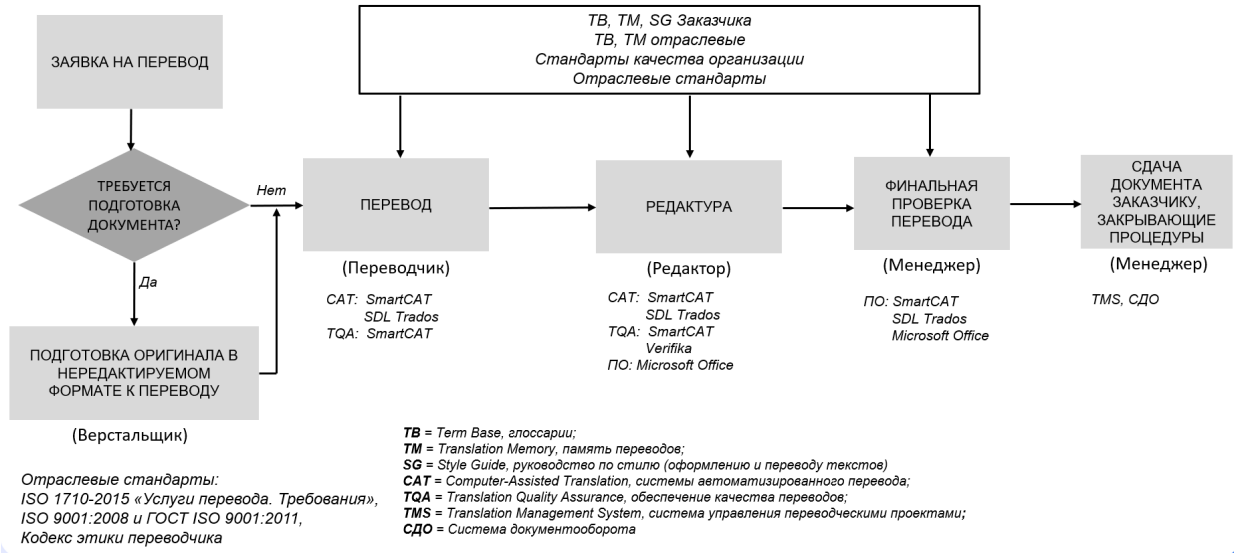

Рисунок 4.8 – Бизнес-процесс осуществления перевода до внедрения программного комплекса

Для оценки производительности переводчиков и редакторов в Агентстве переводов используется оценка количества сданных менеджеру проекта стандартных страниц текста за один час рабочего времени, 1 стандартная страница = 1800 знаков с пробелами. На момент проведения исследования средняя производительность редакторов машинного перевода составила 3,8 страниц в час.

В рамках внедрения программного комплекса в контур автоматизации бизнес-процесса перевода он был адаптирован к специфике деятельности Агентства переводов. Для этого была разработана процедура, позволяющая сотрудникам использовать функциональность программного комплекса без прямой интеграции с системами поддержки переводческой деятельности (CAT).

Процедура включает следующие шаги:

1 Сегментация исходного русскоязычного текста при помощи любой из используемых систем CAT.

2 Экспорт сегментированного текста в файл .xlsx.

3 Использование программного комплекса для получения машинного перевода на английский язык.

4 Сохранение исходного текста и полученного машинного перевода в виде параллельного корпуса в формате таблицы файла. xslx.

5 Импорт полученного корпуса в базу переводов требуемой системы CAT.

6 Редактура полученного перевода в выбранной среде CAT.

В процессе внедрения программного комплекса на стороне Агентства переводов участвовали менеджер проектов и два редактора переводов. Все сотрудники прошли обучение работе с программным комплексом, для чего были разработаны учебные материалы и проведено несколько практических занятий.

Схематично, процесс осуществления перевода в Агентстве переводов до внедрения программного комплекса представлен на рисунке 4.9.

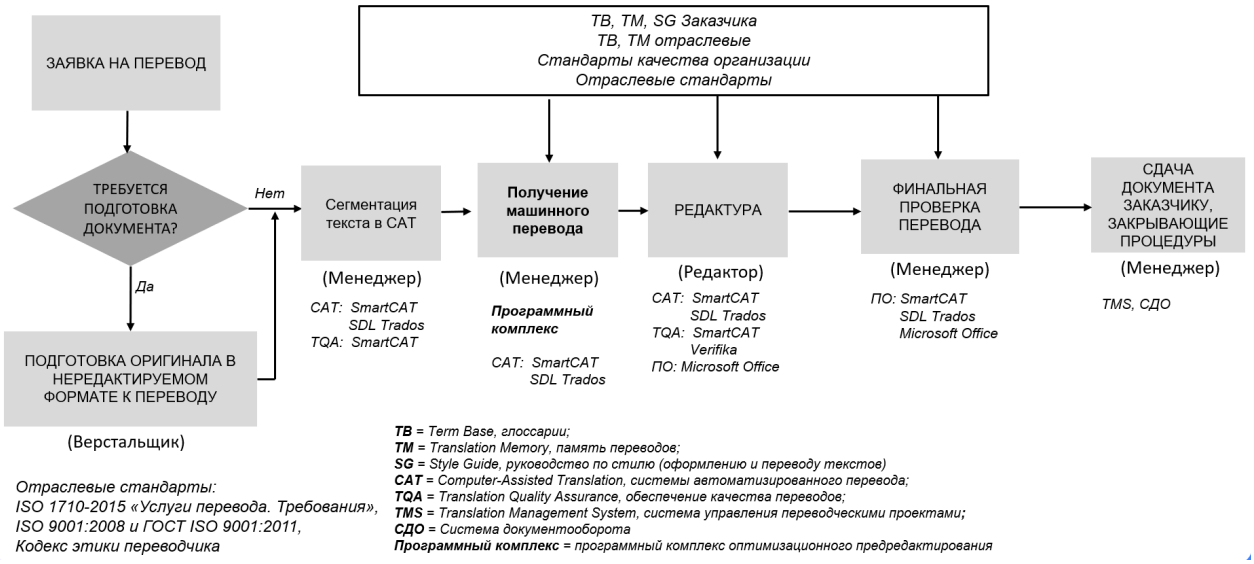

Рисунок 4.9 – Бизнес-процесс осуществления перевода после внедрения программного комплекса

Мониторинг результатов внедрения проводился в течение двух месяцев с момента интеграции программного комплекса в контур автоматизации процессов перевода и включал оценку объемов перевода, производительности редакторов переводов, опрос сотрудников Агентства переводов.

За время мониторинга при помощи программного комплекса было переведено на английский язык 677 стандартных страниц текста. При этом, средняя производительность редакторов переводов при работе с программным комплексом составила 4,3 стандартных страницы в час. Таким образом, внедрение программного комплекса позволило увеличить производительность редакторов переводов на 13,16%.

Опрос сотрудников позволил выявить две основные причины повышения производительности:

1 Повышение качества машинного перевода, ввиду чего сокращается количество требуемых корректировок.

2 Автоматизация работы с проблемными сегментами текста. Использование оценки сложности задачи перевода позволяет отфильтровать сегменты с ожидаемо низким качеством машинного перевода и применить корректирующие действия. Так, менеджер проектов в обратной связи указал, что «удобно производить оценку текста на этапе подготовки к переводу и задать клиенту уточняющие вопросы по спорным моментам, например, по расшифровке аббревиатур».

Для повышения точности оценки сложности задачи перевода и качества машинного перевода рекомендуется систематически проводить дообучение моделей, реализованных в программном комплексе, на данных памятей переводов Агентства переводов.

Выводы по четвертой главе

Результаты тестирования программного комплекса подтверждают работоспособность программного комплекса в части повышения качества машинного перевода узкоспециальных технических текстов путем автоматического оптимизационного предредактирования и использования оценки сложности задачи перевода в качестве в качестве критерия оптимизации.

Программный комплекс в совокупности описанных подсистем и программных модулей имеет потенциал к доработке и повышении точности. Интеграция созданной системы оптимизационного предредактирования в контур автоматизации процессов перевода технической документации позволит снизить затраты на постредактирование МП и организацию переводческих процессов.

Использование полученных результатов исследования и внедрение программного комплекса в работу Агентства переводов позволило повысить эффективность использования машинного перевода и производительность труда редакторов переводов, оптимизировав тем самым затраты на оказание услуг перевода узкоспециальной технической документации.

ЗАКЛЮЧЕНИЕ

В ходе проведения исследования получены следующие основные результаты:

1 Разработана математическая модель процесса перевода и вероятностной оценки сложности задачи перевода, отличительной особенностью которой является возможность получения результатов в аналитическом виде.

2 Предложен новый алгоритм оценки русскоязычного текста по лексическим, синтаксическим и морфологическим признакам, отличающийся возможностью анализа текста по 96 признакам с получением вещественных оценок по каждому из них.

3 Предложен алгоритм определения стратегии оптимизационного предредактирования русскоязычного текста с целью повышения качества его перевода на английский язык по критериям пользователя перевода с использованием моделей машинного перевода, отличающийся тем, что в качестве критерия оптимизации используется вероятностная оценка сложности задачи перевода.

4 Разработана теория вероятностной оценки сложности задачи перевода, позволяющая приближенно вычислять ожидаемое качество перевода заданного текста заданным переводчиком в соответствии с формализованными требованиями к переводу.

5 Предложен новый алгоритм, позволяющий расширить область применения оптимизационного метода градиентного спуска путём использования элементов нечеткой логики в выражении функции правдоподобия через функцию принадлежности полученного текста низкой сложности задачи перевода для выбранной системы МП.

6 Предложен новый алгоритм, позволяющий расширить область применения метода наименьших квадратов для поиска весов значимости параметров исходного текста для вероятностной оценки ожидаемого качества его перевода на целевой язык.

7 Реализован программный комплекс анализа русскоязычного текста и вероятностной оценки сложности задачи его перевода на английский язык с

использованием модели машинного перевода, отличающихся от существующих отсутствием необходимости оптимизации алгоритмов и моделей генерирования текста перевода.

8 Программно реализован алгоритм оценки русскоязычного текста по лексическим, синтаксическим и морфологическим признакам, который позволяет анализировать русскоязычные тексты и определять признаки, значимые при решении задачи перевода с учетом выбранной системы машинного перевода, отличающийся возможностью получения вещественных значений оценок текста.

9 Реализован программный комплекс для повышения качества машинного перевода текстов с русского языка на английский язык, отличающийся от существующих применением оптимизационного предредактирования на основе вероятностной оценки сложности задачи перевода для повышения качества машинного перевода текстов с русского языка на английский язык.

Поставленные цели и задачи исследования достигнуты, что подтверждается результатами тестирования и внедрения разработанных методов, алгоритмов и комплексов программ.

Разработанные методы, алгоритмы и комплексы программ могут быть масштабированы на различные языковые пары и способы перевода, включая ручной перевод, они намечают подходы к управлению рисками, связанными с качеством перевода в зависимости от компетенции выбранных исполнителей, и предоставит индустрии инструменты объективной оценки исполнителей в рамках поставленной задачи на перевод, автоматизированной подготовки текстов к переводу и повышения качества перевода, в том числе для редакторов без знания языка перевода.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1 Search for peer-reviewed journal articles and book chapters [Электронный ресурс] // Science Direct: Platform of Peer-Reviewed Literature. – 2023. – Режим доступа: https://www.sciencedirect.com/

2 Cho K. et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation // arXiv preprint arXiv:1406.1078. – 2014.

3 Castilho S. et al. Is neural machine translation the new state of the art? // The Prague Bulletin of Mathematical Linguistics. – 2017. – №. 108.

4 Wu Y. et al. Google's neural machine translation system: Bridging the gap between human and machine translation // arXiv preprint arXiv:1609.08144. – 2016.

5 Screen B. What effect does post-editing have on the translation product from an end-user’s perspective // The Journal of Specialised Translation. – 2019. – Т. 31. – С. 133-157.

6 Guerberof Arenas A., Moorkens J. Machine translation and post-editing training as part of a master’s programme // Jostrans: The Journal of Specialised Translation. – 2019. – №. 31. – С. 217-238.

7 Кислова Е. Новые смыслы профессии в условиях высокой неопределенности будущего [Электронный ресурс] / Е. Кислова, А. Мустаев, Е.Гайдерова // Пленарное заседание Translation Forum Russia 2023. – 2023. – Режим доступа: https://tconference.ru/provisionalprogramme/

8 Hutchins J. Machine translation: History and general principles // The encyclopedia of languages and linguistics. – 1994. – Т. 5. – С. 2322-2332.

9 Hutchins W. J. Machine translation: A brief history // Concise history of the language sciences. – Pergamon, 1995. – С. 431-445.

10 Hutchins J. Machine translation: history // Encyclopedia of languages and linguistics. – 2006. – С. 375-383.

11 Sreelekha S. et al. A survey report on evolution of machine translation // Int. J. Control Theory Appl. – 2016. – Т. 9. – №. 33. – С. 233-240.

12 Chéragui M. A. Theoretical Overview of Machine translation // ICWIT. – 2012. – С. 160-169.

13 Митренина О. В. Назад, в 47-й: к 70-летию машинного перевода как научного направления //Вестник Новосибирского государственного университета. Серия: Лингвистика и межкультурная коммуникация. – 2017. – Т. 15. – №. 3. – С. 5-12.

14 Schwartz L. The history and promise of machine translation //Innovation and expansion in translation process research. – 2018. – С. 161.

15 Hutchins J. Machine translation: A concise history // Computer aided translation: Theory and practice. – 2007. – Т. 13. – №. 29-70. – С. 11.

16 Holzinger A. From machine learning to explainable AI // 2018 World symposium on digital intelligence for systems and machines (DISA). – IEEE, 2018. – С. 55-66.

17 Al′tshuller G. S. The innovation algorithm: TRIZ, systematic innovation and technical creativity. – Technical innovation center, Inc., 1999.

18 Gadd K. TRIZ for engineers: enabling inventive problem solving. – John Wiley & Sons, 2011.

19 Bertoncelli T., Mayer O., Lynass M. Creativity, learning techniques and TRIZ // Procedia Cirp. – 2016. – Т. 39. – С. 191-196.

20 Berdonosov V. D. Fractality of knowledge and TRIZ // Procedia Engineering. – 2011. – Т. 9. – С. 659-664.

21 Altshuller G. 40 principles: TRIZ keys to technical innovation. – Technical Innovation Center, Inc., 2002. – Т. 1.

22 Бердоносов В. Д., Животова А. А. Исследование эволюции объектно-ориентированных языков программирования // Ученые записки Комсомольского-на-Амуре государственного технического университета. – 2014. – Т. 1. – №. 2. – С. 35-43.

23 Berdonosov V. et al. Perspectives for development of automatized enterprise management systems // Proceedings of the TRIZfest-2015 International Conference. – 2015. – С. 154-163.

24 Berdonosov V. D., Kozlita A. N., Zhivotova A. A. TRIZ evolution of black oil coker units // Chemical engineering research and design. – 2015. – Т. 103. – С. 61-73.

25 Berdonosov V. D., Redkolis E. V. TRIZ evolutionary approach: Main points and implementation // Research and Practice on the Theory of Inventive Problem Solving (TRIZ) Linking Creativity, Engineering and Innovation. – 2016. – С. 95-111.

26 Hutchins J., Lovtskii E. Petr Petrovich Troyanskii (1894–1950): A forgotten pioneer of mechanical translation // Machine translation. – 2000. – Т. 15. – С. 187-221.

27 Троянский, П. П. Авторское свидетельство № 40995 А1 СССР, МПК В41В 13/00. Машина для подбора и печатания слов при переводе с одного языка на другой : № 134430 : заявл. 05.09.1933 : опубл. 31.01.1935.

28 Oswald V. A. Word-by-word translation // Proceedings of the Conference on Mechanical Translation. – 1952.

29 Yngve V. H., Charney E. K., Klima E. S. Mechanical translation. – Research Laboratory of Electronics (RLE) at the Massachusetts Institute of Technology (MIT), 1962.

30 Booth A. D. The practical realisation of machine translation // Methodos. – 1956. – Т. 8. – С. 23-33.

31 Brown A. F. R. Automatic translation of languages // Information storage and Retrieval. – 1964. – Т. 2. – №. 1. – С. 1-28.

32 Tosh L. W. Data preparation for syntactic translation // COLING 1965. – 1965.

33 Masterman M. Semantic message detection for machine translation, using an interlingua //Proceedings of the International Conference on Machine Translation and Applied Language Analysis. – 1961. – С. 438-474.

34 Tosh L. Stratificational grammar and interlingual mapping for automatic translation // Actes du Xe Congrès International des Linguistes. – 1967. – Т. 4. – С. 1049-1059.

35 Aramaki E., Kurohashi S. Example-based machine translation using structural translation examples // Proceedings of the First International Workshop on Spoken Language Translation: Evaluation Campaign. – 2004. – С.91-94.

36 Nepveu L. et al. Adaptive language and translation models for interactive machine translation // Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing. – 2004. – С. 190-197.

37 Brown P. F. et al. A statistical approach to machine translation // Computational linguistics. – 1990. – Т. 16. – №. 2. – С. 79-85.

38 Junczys-Dowmunt M., Grundkiewicz R. Phrase-based machine translation is state-of-the-art for automatic grammatical error correction // arXiv preprint arXiv:1605.06353. – 2016.

39 Yamada K. A syntax-based statistical translation model. – University of Southern California, 2003.

40 Johnson M. et al. Google’s multilingual neural machine translation system: Enabling zero-shot translation // Transactions of the Association for Computational Linguistics. – 2017. – Т. 5. – С. 339-351.

41 Lample G. et al. Unsupervised machine translation using monolingual corpora only // arXiv preprint arXiv:1711.00043. – 2017.

42 Chu C., Wang R. A survey of domain adaptation for neural machine translation // arXiv preprint arXiv:1806.00258. – 2018.

43 Costa-Jussa M. R., Fonollosa J. A. R. Latest trends in hybrid machine translation and its applications // Computer Speech & Language. – 2015. – Т. 32. – №. 1. – С. 3-10.

44 Животова А. А., Бердоносов В. Д. Перспективные направления развития систем машинного перевода // Информатика и системы управления. – 2022. – №2(72). – С.116-132.

45 Zhivotova A. A., Berdonosov V. D., Redkolis E. V. Machine translation systems analysis and development prospects // Proceedings of 2020 International Multi-Conference on Industrial Engineering and Modern Technologies (FarEastCon-2020). – Vladivostok: Russia. 2020.

46 Doherty S., O'Brien S. Assessing the usability of raw machine translated output: A user-centered study using eye tracking // International Journal of Human-Computer Interaction. – 2014. – Т. 30. – №. 1. – С. 40-51.

47 Lear A. et al. Why Can't I Just Use Google Translate? A Study on the Effectiveness of Online Translation Tools in Translation of Coas // Value in Health. – 2016. – Т. 19. – №. 7. – С. A387.

48 Quinci C., Pontrandolfo G. Testing neural machine translation against different levels of specialization // trans-com. – 2023. – №16. – С.174-209.

49 Canfora C., Ottmann A. Risks in neural machine translation // Translation Spaces. – 2020. – Т. 9. – №. 1. – С. 58-77.

50 Животова, А. А. Автоматизация предредактирования исходного текста для повышения качества машинного перевода / А. А. Животова, В. Д. Бердоносов, И. А. Лошманова // Наука, инновации и технологии: от идей к внедрению : Материалы II Международной научно-практической конферен-ции молодых ученых, Комсомольск-на-Амуре, 14–18 ноября 2022 года / Редколлегия: А.В. Космынин (отв. ред.) [и др.]. Том 1. – Комсомольск-на-Амуре: Комсомольский-на-Амуре государственный университет, 2022. – С. 366-370.

51 Kumar V. et al. A machine assisted human translation system for technical documents // Proceedings of the 8th International Conference on Knowledge Capture. – 2015. – С. 1-5.

52 Azadi F., Khadivi S. Improved search strategy for interactive predictions in computer-assisted translation // Proceedings of Machine Translation Summit XV: Papers. – 2015. – Т.1. – С.319–332.

53 Federico M. et al. Machine translation enhanced computer assisted translation // Nice MT Summit XIV. – 2013. – Т. 1. – С. 425.

54 Alabau V. et al. User evaluation of advanced interaction features for a computer-assisted translation workbench // Machine Translation Summit XIV. – 2013. – Т.1. – С.361–368.

55 Yamada M. The impact of Google neural machine translation on post-editing by student translators // The Journal of Specialised Translation. – 2019. – Т. 31. – №. 1. – С. 87-106.

56 Balling L. W., Carl M., O'Brian S. (ed.). Post-editing of machine translation: Processes and applications. – Cambridge Scholars Publishing, 2014.

57 Herbig N. et al. Multi-modal approaches for post-editing machine translation // Proceedings of the 2019 CHI conference on human factors in computing systems. – 2019. – С. 1-11.

58 Toledo Báez M. C. Machine translation and post-editing: impact of training and directionality on quality and productivity // Tradumàtica. – 2018. – №. 16. – С. 24-34.

59 Feifei F. et al. A Study of Pre-editing Methods at the Lexical Level in the Process of Machine Translation // Arab World English Journal for Translation & Literary Studies. – 2022. – T.6. – № 2. – C.54-69.

60 Zheng Y., Peng C., Mu Y. Designing Controlled Chinese Rules for MT Pre-Editing of Product Description Text //International Journal of Translation, Interpretation, and Applied Linguistics (IJTIAL). – 2022. – T. 4. – №. 2. – C. 1-13.

61 Seretan V., Bouillon P., Gerlach J. A Large-Scale Evaluation of Pre-editing Strategies for Improving User-Generated Content Translation // LREC. – 2014. – C. 1793-1799.

62 Shei C. C. Teaching MT through pre-editing: Three case studies // Proceedings of the 6th EAMT Workshop: Teaching Machine Translation. – 2002.

63 Liang Y., Han W. Source text pre-editing versus target text post-editing in using Google Translate to provide health services to culturally and linguistically diverse clients // Science, Engineering and Health Studies. – 2022. – T. 16. – C. 22050009-22050009.

64 Gerlach J. et al. Combining pre-editing and post-editing to improve SMT of user-generated content // Proceedings of the 2nd Workshop on Post-editing Technology and Practice. – 2013.

65 Yu K. N. et al. Pre-editing English news texts for machine translation into Russian // Language Studies and Modern Humanities. – 2022. – T. 4. – №. 1. – C. 25-30.

66 Marzouk S., Hansen-Schirra S. Evaluation of the impact of controlled language on neural machine translation compared to other MT architectures // Machine Translation. – 2019. – T. 33. – №. 1-2. – C. 179-203.

67 Arenas A. G. Pre-editing and post-editing // The Bloomsbury companion to language industry studies. – 2019. – C. 333-360.

68 Sánchez-Gijón P., Kenny D. Selecting and preparing texts for machine translation: Pre-editing and writing for a global audience // Machine translation for everyone: Empowering users in the age of artificial intelligence. – 2022. – T. 18. – C. 81.

69 Zhivotova A. A., Berdonosov V. D., Redkolis E. V. Improving the Quality of Scientific Articles Machine Translation while Writing Original Text // Proceedings of 2020 International Multi-Conference on Industrial Engineering and Modern Technologies (FarEastCon-2020). – Vladivostok: Russia, 2020.

70 Machine translation tips [Электронный ресурс] // IBM Cloud Docs. – 2016. Режим доступа: https://cloud.ibm.com/docs/GlobalizationPipeline?topic=GlobalizationPipeline-globalizationpipeline_tips

71 Writing for a global audience [Электронный ресурс] // Google developer documentation style guide. – 2020. Режим доступа: https://developers.google.com/style/translation.

72 Miyata R. et al. Evaluating the usability of a controlled language authoring assistant // The Prague bulletin of mathematical linguistics. – 2017. – Т. 108. – №. 1. – С. 147.

73 O'Brien S. Controlling controlled english // EAMT Workshop: Improving MT through other language technology tools: resources and tools for building MT. – 2003.

74 Осокина С. А. Концепция «легкого языка» и перспективы ее развития в лингвистике // Филология и человек. – 2022. – №. 2. – С. 115-133.

75 Muegge U. Controlled Language Optimized for Uniform Translation (CLOUT). – Bepress, 2002.

76 Polo L. R. Controlled Language and the Implementation of Machine Translation for Technical Documentation // Translating and the Computer 27. – 2005.

77 Hiraoka Y., Yamada M. Pre-editing plus neural machine translation for subtitling: effective pre-editing rules for subtitling of TED Talks // Proceedings of Machine Translation Summit XVII: Translator, Project and User Tracks. – 2019. – С. 64-72.

78 Miyata R., Fujita A. Dissecting human pre-editing toward better use of off-the-shelf machine translation systems // Proceedings of the 20th annual conference of the european association for machine translation (EAMT). – 2017. – С. 54-59.

79 Taufik A. Pre-editing of Google neural machine translation //Journal of English Language and Culture. – 2020. – Т. 10. – №. 2. – С. 64–74.

80 Mercader-Alarcón J., Sánchez-Matínez F. Analysis of translation errors and evaluation of pre-editing rules for the translation of English news texts into Spanish with Lucy LT //Tradumàtica: traducció i tecnologies de la informació i la comunicació. – 2016. – №. 14. – С. 172-186.

81 Ниценко А. В., Шелепов В. Ю. О некоторых подходах к проблеме автоматической адаптации русскоязычных текстов //Программная инженерия: методы и

технологии разработки информационно-вычислительных систем (ПИИВС-2020). – 2020. – С. 77-83.

82 Li Z. et al. Paraphrase generation with deep reinforcement learning //arXiv preprint arXiv:1711.00279. – 2017.

83 Zhou J., Bhat S. Paraphrase generation: A survey of the state of the art //Proceedings of the 2021 conference on empirical methods in natural language processing. – 2021. – С. 5075-5086.

84 Aluisio S. et al. Readability assessment for text simplification //Proceedings of the NAACL HLT 2010 fifth workshop on innovative use of NLP for building educational applications. – 2010. – С. 1-9.

85 Siddharthan A. A survey of research on text simplification //ITL-International Journal of Applied Linguistics. – 2014. – Т. 165. – №. 2. – С. 259-298.

86 Drndarevic B., Saggion H. Reducing text complexity through automatic lexical simplification: an empirical study for Spanish //Procesamiento del lenguaje natural. – 2012. – Т. 49. – С. 13-20.

87 Сибирцева В.Г., Карпов Н.В. Автоматическая адаптация текстов для электронных учебников. Проблемы и перспективы (на примере русского языка) // Новая русистика. – 2014. – №7. – С.19-33.

88 Русское синтаксическое управление при словесных заменах. О словах с функциями наречия и существительного / А. В. Ниценко, В. Ю. Шелепов, С. А. Большакова, К. С. Ивашко // Проблемы искусственного интеллекта. – 2020. – № 2(17). – С. 46-57.

89 О словесных заменах, сохраняющих смысл русского предложения / А. В. Ниценко, В. Ю. Шелепов, С. А. Большакова, К. С. Ивашко // Проблемы искусственного интеллекта. – 2020. – № 1(16). – С. 63-74.

90 Fabre B. et al. Neural-Driven Search-Based Paraphrase Generation //Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume. – 2021. – С. 2100-2111.

91 Sokolov A., Filimonov D. Neural machine translation for paraphrase generation //arXiv preprint arXiv:2006.14223. – 2020.

92 Оборнева, И. В. Автоматизация оценки качества восприятия текста // ВЕСТНИК Московского городского педагогического университета, 2015. – №2(5). – С. 221–233.

93 Дмитриева А. Д., Лапошина А. Н., Лебедева М. Ю. Квантитативное исследование стратегий упрощения на материале адаптированных текстов для изучающих РКИ //Компьютерная лингвистика и интеллектуальные технологии: по материалам международной конференции «Диалог». М.: РГГУ. – 2021. – С. 191-204.

94 Church K., Liberman M. The future of computational linguistics: On beyond alchemy //Frontiers in Artificial Intelligence. – 2021. – Т. 4. – С. 625341.

95 Moore J. D., Wiemer-Hastings P. Discourse in computational linguistics and artificial intelligence //Handbook of discourse processes. – 2003. – С. 439-486.

96 Zhivotova A. A., Berdonosov V. D., Gordin S. A. Mathematical Modeling of the Translation Process and Its Optimization by the Criterion of Quality Maximization // Information Technologies and Intelligent Decision-Making Systems: Communications in Computer and Information Science. – 2023. – vol. 1821. – P. 1–15.

97 Чернявская, В. Е. Лингвистика текста. Лингвистика дискурса / В. Е. Чернявская. – М.: Общество с ограниченной ответственностью «ФЛИНТА», 2012. – 208 с.

98 Мизернов И. Ю., Гращенко Л. А. Анализ методов оценки сложности текста // Новые информационные технологии в автоматизированных системах. – 2015. – №. 18. – С. 572-581.

99 Солнышкина М. И., Казачкова М. Б., Харькова Е. В. Инструменты измерения сложности текстов на английском языке // Иностр. языки в школе. – 2020. – №. 3. – С. 15.

100 Hosmer D. W., Lemeshow S. Applied Logistic Regression 2nd edn Wiley & Sons //New York. – 2000.

101 Животова А. А., Бердоносов В. Д. Стратегия предредактирования исходного текста на основании автоматической оценки сложности задачи перевода для повышения качества машинного перевода узкоспециальных текстов на английский язык // Компьютерная лингвистика и интеллектуальные технологии: по материалам

ежегодной международной конференции «Диалог». – № 22. – Доп. том. – 2023. – С. 1141-1149.

102 Каханер Д., Моулер К., Нэш С. Численные методы и программное обеспечение. – М.: Мир. – 1998.

103 MacKay D. J. C. Information theory, inference and learning algorithms. – Cambridge university press, 2003.

104 Barber D. Bayesian reasoning and machine learning. – Cambridge University Press, 2012.

105 Bishop C. Pattern recognition and machine learning //Springer google schola. – 2006. – Т. 2. – С. 531-537.

106 Bengio Y. Practical recommendations for gradient-based training of deep architectures //Neural Networks: Tricks of the Trade: Second Edition. – Berlin, Heidelberg: Springer Berlin Heidelberg, 2012. – С. 437-478.

107 Raffel C. et al. Exploring the limits of transfer learning with a unified text-to-text transformer //The Journal of Machine Learning Research. – 2020. – Т. 21. – №. 1. – С. 5485-5551.

108 Васильев Д. Д., Пятаева А. В. Использование языковых моделей T5 для задачи упрощения текста //Программные продукты и системы. – 2023. – Т. 36. – №. 2. – С. 228-236.

109 Комиссаров В. Н., Цвиллинг М. Я. Лингвистика перевода. Изд. 5, стереотип. – М.: URSS, 2020. – 176с.

110 Животова А. А., Бердоносов В. Д. Автоматизация предредактирования русскоязычных текстов с целью повышения качества их машинного перевода на английский язык // Информационные технологии и высокопроизводительные вычисления: материалы VII Международной науч.- практ. конф., Хабаровск, 11-13 сентября 2023 г. / Редколлегия: Р.В. Намм (отв. редактор) [и др.]. – ХФИЦ ДВО РАН: Хабаровск. – 2023. – С. 88-91.

111 Nivre J. et al. Universal Dependencies v2: An evergrowing multilingual treebank collection //arXiv preprint arXiv:2004.10643. – 2020.

112 Животова А. А., Бердоносов В. Д. Машинный перевод корпусов текста для прикладных и исследовательских задач // Молодежь и наука: актуальные проблемы фундаментальных и прикладных исследований: Материалы VI Всероссийской национальной научной конференции молодых учёных, Комсомольск-на-Амуре, 10–14 апреля 2023 года. Том 2. – Комсомольск-на-Амуре: Комсомольский-на-Амуре государственный университет, 2023. – С. 467-470.

113 Пикалев Я. С. Обзор архитектур систем интеллектуальной обработки естественно-языковых текстов // Проблемы искусственного интеллекта. – 2020. – № 4(19). – С. 45-68.

114 Findings of the 2019 Conference on Machine Translation (WMT19) / O. Bojar, M. R. Costajuss et al. // Proceedings of the Fourth Conference on Machine Translation. – Florence, Italy: Association for Computational Linguistics. – 2019. – P. 1-61.

115 OPUS-MT – Building open translation services for the World / J. Tiedemann, S. Thottingal // Proceedings of the 22nd Annual Conference of the European Association for Machine Translation. – Lisboa, Portugal: European Association for Machine Translation. – 2020.

116 Han A. L. F. et al. Language-independent model for machine translation evaluation with reinforced factors //Proceedings of Machine Translation Summit XIV: Posters. – 2013. – С. 215-222.

117 Животова А. А., Бердоносов В. Д. Автоматизированная оценка параметров русскоязычного текста // Молодежь и наука: актуальные проблемы фундаментальных и прикладных исследований: Материалы VI Всероссийской национальной научной конференции молодых учёных, Комсомольск-на-Амуре, 10–14 апреля 2023 года. Том 2. – Комсомольск-на-Амуре: Комсомольский-на-Амуре государственный университет, 2023. – С. 464-467.

118 Saggion H. Automatic text simplification: Synthesis lectures on human language technologies, vol. 10 (1) //California, Morgan & Claypool Publishers. – 2017. – С. 137.

119 Частотный словарь современного русского языка (на материалах Национального корпуса русского языка) / О. Н. Ляшевская, С. А. Шаров. – М.: Азбуковник, 2009.

120 Проект Natasha — набор Python-библиотек для обработки текстов на естественном русском языке [Электронный ресурс] // Alexander Kukushkin Data Science Laboratory. – 2023. Режим доступа: https://natasha.github.io/.

121 Животова, А. А. Практическое применение вероятностной оценки сложности задачи перевода // Молодежь и наука: актуальные проблемы фундаментальных и прикладных исследований: Материалы VI Всероссийской национальной научной конференции молодых учёных, Комсомольск-на-Амуре, 10–14 апреля 2023 года. Том 2. – Комсомольск-на-Амуре: Комсомольский-на-Амуре государственный университет, 2023. – С. 461-464.

122 Животова А. А., Бердоносов В. Д., Регрессионный анализ корреляции качества машинного перевода и параметров исходного текста // Информатика и системы управления. – 2023. – №2(76). – С.121-133.

123 Бенгфорт Б., Билбро Р., Охеда Т. Прикладной анализ текстовых данных на Python. Машинное обучение и создание приложений обработки естественного языка. — Санкт-Петербург: Питер, 2019.

124 Животова А.А., Бердоносов В.Д. Оптимизационное предредактирование узкоспециальных русскоязычных текстов для их машинного перевода на английский язык // Информационные и математические технологии в науке и управлении. – 2024. – № 2(34). *(в издании)*

ПРИЛОЖЕНИЕ А

Охранные документы на результаты интеллектуальной деятельности

РОССИЙСКАЯ ФЕДЕРАЦИЯ

СВИДЕТЕЛЬСТВО

о государственной регистрации программы для ЭВМ

№ 2023682260

Программный комплекс для предредактирования и машинного перевода узкоспециальных русскоязычных текстов на английский язык

Правообладатель: ***Животова Алена Анатольевна (RU)***

Авторы: ***Животова Алена Анатольевна (RU), Бердоносов Виктор Дмитриевич (RU)***

Заявка № **2023680875**

Дата поступления **09 октября 2023 г.**

Дата государственной регистрации

в Реестре программ для ЭВМ ***24 октября 2023 г.***

Руководитель Федеральной службы по интеллектуальной собственности

ДОКУМЕНТ ПОДПИСАН ЭЛЕКТРОННОЙ ПОДПИСЬЮ

Сертификат 429b6a0[illegible]164ba96f83b73b4aa7

Владелец Зубов Юрий Сергеевич

Действителен с [illegible]2023 по 02.08.2024

Ю.С. Зубов

РОССИЙСКАЯ ФЕДЕРАЦИЯ

СВИДЕТЕЛЬСТВО

о государственной регистрации программы для ЭВМ

№ 2023669254

Программа для тренировки языковой модели в решении задач перефразирования русскоязычных технических текстов

Правообладатель: ***Животова Алена Анатольевна (RU)***

Автор(ы): ***Животова Алена Анатольевна (RU)***

Заявка № **2023667648**
Дата поступления **25 августа 2023 г.**
Дата государственной регистрации
в Реестре программ для ЭВМ ***12 сентября 2023 г.***

Руководитель Федеральной службы
по интеллектуальной собственности

Ю.С. Зубов

РОССИЙСКАЯ ФЕДЕРАЦИЯ

СВИДЕТЕЛЬСТВО

о государственной регистрации базы данных

№ 2023623048

Корпус параллельных двуязычных текстов нефтегазовой тематики для тренировки языковых моделей в задачах перефразирования узкоспециальных технических русскоязычных текстов и повышения качества их перевода на английский язык

Правообладатель: ***Животова Алена Анатольевна (RU)***

Автор(ы): ***Животова Алена Анатольевна (RU)***

Заявка № **2023622757**
Дата поступления **25 августа 2023 г.**
Дата государственной регистрации
в Реестре баз данных ***06 сентября 2023 г.***

Руководитель Федеральной службы по интеллектуальной собственности

Ю.С. Зубов

РОССИЙСКАЯ ФЕДЕРАЦИЯ

СВИДЕТЕЛЬСТВО

о государственной регистрации базы данных

№ 2023623022

База данных структурного анализа предложений технических русскоязычных текстов

Правообладатель: ***Животова Алена Анатольевна (RU)***

Автор(ы): ***Животова Алена Анатольевна (RU)***

Заявка № **2023622765**
Дата поступления **25 августа 2023 г.**
Дата государственной регистрации
в Реестре баз данных ***01 сентября 2023 г.***

Руководитель Федеральной службы
по интеллектуальной собственности

Ю.С. Зубов

РОССИЙСКАЯ ФЕДЕРАЦИЯ

СВИДЕТЕЛЬСТВО

о государственной регистрации программы для ЭВМ

№ 2023668511

Предредактор русскоязычных узкоспециальных текстов для систем машинного перевода

Правообладатель: ***Животова Алена Анатольевна (RU)***

Автор(ы): ***Животова Алена Анатольевна (RU)***

Заявка № **2023667667**
Дата поступления **25 августа 2023 г.**
Дата государственной регистрации
в Реестре программ для ЭВМ ***29 августа 2023 г.***

Руководитель Федеральной службы по интеллектуальной собственности

Ю.С. Зубов

РОССИЙСКАЯ ФЕДЕРАЦИЯ

СВИДЕТЕЛЬСТВО

о государственной регистрации программы для ЭВМ

№ 2023665348

Программный модуль автоматической очистки сырых данных из памятей переводов CAT для тренировки моделей нейронного машинного перевода

Правообладатель: ***Животова Алена Анатольевна (RU)***

Автор(ы): ***Животова Алена Анатольевна (RU)***

Заявка № **2023664365**
Дата поступления **10 июля 2023 г.**
Дата государственной регистрации
в Реестре программ для ЭВМ ***14 июля 2023 г.***

Руководитель Федеральной службы по интеллектуальной собственности

Ю.С. Зубов

РОССИЙСКАЯ ФЕДЕРАЦИЯ

СВИДЕТЕЛЬСТВО

о государственной регистрации программы для ЭВМ

№ 2023663773

ПРОГРАММНЫЙ КОМПЛЕКС ДЛЯ АНАЛИЗА РУССКОЯЗЫЧНОГО ТЕКСТА И ВЕРОЯТНОСТНОЙ ОЦЕНКИ СЛОЖНОСТИ ЗАДАЧИ ЕГО ПЕРЕВОДА НА АНГЛИЙСКИЙ ЯЗЫК С ИСПОЛЬЗОВАНИЕМ МОДЕЛИ МАШИННОГО ПЕРЕВОДА

Правообладатель: *Животова Алена Анатольевна (RU)*

Автор(ы): *Животова Алена Анатольевна (RU)*

Заявка № **2023663302**
Дата поступления **27 июня 2023 г.**
Дата государственной регистрации
в Реестре программ для ЭВМ *28 июня 2023 г.*

Руководитель Федеральной службы по интеллектуальной собственности

Ю.С. Зубов

РОССИЙСКАЯ ФЕДЕРАЦИЯ

RU **2023617748**

ФЕДЕРАЛЬНАЯ СЛУЖБА
ПО ИНТЕЛЛЕКТУАЛЬНОЙ СОБСТВЕННОСТИ

(12) **ГОСУДАРСТВЕННАЯ РЕГИСТРАЦИЯ ПРОГРАММЫ ДЛЯ ЭВМ**

Номер регистрации (свидетельства):
2023617748

Дата регистрации: **13.04.2023**

Номер и дата поступления заявки:
2023613934 28.02.2023

Дата публикации: **13.04.2023**

Контактные реквизиты:
patent@knastu.ru

Авторы:
Животова Алена Анатольевна (RU),
Бердоносов Виктор Дмитриевич (RU)

Правообладатель:
Федеральное государственное бюджетное образовательное учреждение высшего образования «Комсомольский-на-Амуре государственный университет» (ФГБОУ ВО «КнАГУ») (RU)

Название программы для ЭВМ:
«ПРОГРАММА ДЛЯ ВЕРОЯТНОСТНОЙ ОЦЕНКИ СЛОЖНОСТИ ПЕРЕВОДЧЕСКОЙ ЗАДАЧИ ДЛЯ СИСТЕМ МАШИННОГО ПЕРЕВОДА»

Реферат:
Программа предназначена для оценки сложности переводческой задачи, обратно пропорциональной вероятности получения перевода, который соответствует требованиям качества пользователя, при его выполнении системой машинного перевода. Программа обучается на массиве данных, включающих исходный текст, эталонный перевод и перевод, выполненный системой машинного перевода. В качестве математической модели оценки вероятности используется модель логистической регрессии. Обученная модель решает задачу классификации, рассчитывая вероятность попадания переведенного текста в класс «качественный перевод» в соответствии с требованиями пользователя. Область применения программы: прикладная компьютерная лингвистика, задачи повышения качества машинного перевода. Тип ЭВМ: IBM PC - совмест. ПК; ОС: Windows, Mac OS X, Linux.

Язык программирования: Python

Объем программы для ЭВМ: 570 КБ

РОССИЙСКАЯ ФЕДЕРАЦИЯ

СВИДЕТЕЛЬСТВО

о государственной регистрации программы для ЭВМ

№ 2023614410

ПРЕПРОЦЕССИНГ ТЕКСТОВЫХ ДАННЫХ ДЛЯ ВЗВЕШЕННОЙ ОЦЕНКИ ПАРАМЕТРОВ РУССКОЯЗЫЧНОГО ТЕКСТА

Правообладатель: ***Животова Алена Анатольевна (RU)***

Автор(ы): ***Животова Алена Анатольевна (RU)***

Заявка № **2023612552**
Дата поступления **13 февраля 2023 г.**
Дата государственной регистрации
в Реестре программ для ЭВМ ***01 марта 2023 г.***

Руководитель Федеральной службы по интеллектуальной собственности

Ю.С. Зубов

РОССИЙСКАЯ ФЕДЕРАЦИЯ

СВИДЕТЕЛЬСТВО

о государственной регистрации программы для ЭВМ

№ 2023613906

ПАРСЕР МАШИННОГО ПЕРЕВОДА УЗКОСПЕЦИАЛЬНЫХ ТЕХНИЧЕСКИХ ТЕКСТОВ С РУССКОГО ЯЗЫКА НА АНГЛИЙСКИЙ ЯЗЫК

Правообладатель: *Животова Алена Анатольевна (RU)*

Автор(ы): *Животова Алена Анатольевна (RU)*

Заявка № **2023612511**
Дата поступления **13 февраля 2023 г.**
Дата государственной регистрации
в Реестре программ для ЭВМ *21 февраля 2023 г.*

Руководитель Федеральной службы по интеллектуальной собственности

документ подписан электронной подписью
Сертификат [illegible]
Владелец Зубов Юрий Сергеевич
Действителен с [illegible] по 26.05.2023

Ю.С. Зубов

ПРИЛОЖЕНИЕ Б

Акт о внедрении (использовании) результатов кандидатской диссертационной работы

ООО «Агентство переводов «ФИАС-Амур»
681013, Россия, Хабаровский край, г. Комсомольск-на-Амуре, пр. Ленина, 34/2, оф. 15
Тел./Факс: +7 (4217) 244-344, +7 (4217) 55-37-25, e-mail: translate@fiasamur.ru, www.fiasamur.ru
Translation Agency FIAS-Amur Co. LTD
34/2, Lenin Avenue, Suite 15, Komsomolsk-on-Amur, Khabarovsky Kray, Russia, 681013
Phone/Fax: +7 (4217) 244-344, +7 (4217) 55-37-25, e-mail: translate@fiasamur.ru, www.fiasamur.ru

УТВЕРЖДАЮ
Генеральный директор
ООО «Агентство переводов «ФИАС-Амур»
К.Н. Войтик
6831 от «10» июня 2023 г.

АКТ

о внедрении (использовании) результатов
кандидатской диссертационной работы
Животовой Алены Анатольевны

Комиссия в составе: председатель Кира Николаевна Войтик, члены комиссии: Галина Юрьевна Мастевная, Сергей Петрович Светлаков, составили настоящий акт о том, что результаты диссертационной работы Животовой Алены Анатольевны «Математическая модель, алгоритмы и программный комплекс для повышения качества машинного перевода узкоспециальных технических текстов на английский язык», представленной на соискание ученой степени кандидата технических наук, использованы в исследовательской и практической деятельности ООО «Агентство переводов «ФИАС-Амур» при разработке контура автоматизации переводческой деятельности и создания подсистемы поддержки принятия решений для менеджеров проектов и управления рисками, связанными с компетенцией исполнителей, в том числе при использовании систем машинного перевода, а также создания баз данных для тренировки языковых моделей.

В ходе исследования Животовой Аленой Анатольевной были разработаны методика и алгоритм вероятностной оценки сложности задачи перевода, методика оптимизационного предредактирования русскоязычных текстов. Разработанные алгоритмы реализованы в программном комплексе для оценки сложности задачи перевода русскоязычных текстов на английский язык и оптимизационного предредактирования и прошли валидацию на корпусах двуязычных текстов (языковые пары русский-английский и английский-русский) объемом – 40 000 страниц, собранных на базе Translation Memory из сред автоматизации перевода, применяемых ООО «Агентство переводов «ФИАС-Амур», получены свидетельства о регистрации программ ЭВМ и баз данных.

Заключение: Использование полученных результатов исследования и внедрение программного комплекса в работу агентства переводов позволило повысить эффективность использования систем машинного перевода и производительность труда редакторов переводов, а также оптимизировать затраты на оказание услуг перевода узкоспециальной технической документации.

Председатель комиссии:
Генеральный директор — Войтик К.Н.

Члены комиссии:
Главный бухгалтер — Мастевная Г.Ю.
Руководитель службы автоматизации — Светлаков С.П.