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

АВТОРЕФЕРАТ

диссертации на соискание учёной степени кандидата технических наук

Комсомольск-на-Амуре – 2024

Работа выполнена в федеральном государственном бюджетном образовательном учреждении высшего образования «Комсомольский-на-Амуре государственный университет», г. Комсомольск-на-Амуре; Вычислительном центре Дальневосточного отделения Российской академии наук – обособленном подразделении федерального государственного бюджетного учреждения науки «Хабаровский Федеральный исследовательский центр Дальневосточного отделения Российской академии наук», г. Хабаровск.

Научный руководитель: **БЕРДОНОСОВ Виктор Дмитриевич,** кандидат технических наук, доцент, доцент кафедры «Прикладная математика» ФГБОУ ВО «Комсомольский-на-Амуре государственный университет», г. Комсомольск-на-Амуре

Официальные оппоненты: **АРТЕМЬЕВА Ирина Леонидовна,** доктор технических наук, профессор, заместитель директора по науке Института математики и компьютерных технологий ФГАОУ ВО «Дальневосточный федеральный университет», г. Владивосток

КОЖЕМЯКИНА Ольга Юрьевна, доктор технических наук, канд. филол. наук, ведущий научный сотрудник ФГБНУ «Федеральный исследовательский центр информационных и вычислительных технологий Сибирского отделения Российской академии наук», г. Новосибирск

Ведущая организация: Федеральное государственное бюджетное образовательное учреждение высшего образования «Санкт-Петербургский государственный университет», г. Санкт-Петербург

Защита состоится 17 октября 2024 г. в 15:00 на заседании диссертационного совета Д 24.1.478.02, созданного на базе Хабаровского Федерального исследовательского центра Дальневосточного отделения Российской академии наук, по адресу: 680000 г. Хабаровск, ул. Ким-Ю-Чена 65, ауд. 118.

С диссертацией можно ознакомиться в научной библиотеке ХФИЦ ДВО РАН в обособленном подразделении «Вычислительный центр ДВО РАН» по адресу: 680000, г. Хабаровск, ул. Ким-Ю-Чена 65 и в библиотеке ФГБОУ ВО «Комсомольский-на-Амуре государственный университет» по адресу: 681013, г. Комсомольск-на-Амуре, пр. Ленина, 27.

Отзывы на автореферат в двух экземплярах, заверенные печатью учреждения, просим направлять по адресу: 680000, г. Хабаровск, ул. Ким-Ю-Чена 65, ученому секретарю диссертационного совета Д 24.1.478.02.

Автореферат разослан «___»_____2024 г.
Телефон для справок: +7 (4212) 32-79-27.

Учёный секретарь
диссертационного совета
Д 24.1.478.02,
канд. техн. наук, доцент  Пассар Андрей Владимирович

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность темы исследования. 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 машинного перевода. 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.

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 незначительно освещена, однако известны работы И. В. Оборневой, А. Н. Лапошиной, М. Ю. Лебедевой 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 соответствует 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 внедрения результатов диссертации.

ОСНОВНОЕ СОДЕРЖАНИЕ РАБОТЫ

Во введении дано обоснование актуальности и характеристика работы, сформулированы цель и задачи, указаны методы исследования, представлены основные положения, показана научная новизна и практическая значимость, описана структура работы.

В первой главе предложено использовать ТРИЗ-эволюционный подход к выявлению направлений развития гибридных систем нейронного машинного перевода (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.

Анализ существующих исследований по теме автоматизированной предобработки исходных текстов для МП показал, что:

- отсутствуют исследования для перевода текстов с русского языка, хотя для русского языка разработаны синтаксические анализаторы и возможна разработка правил предредактирования для повышения качества МП;
- в литературе не описаны правила и методологии предредактирования исходных текстов для перевода в системах нейронного МП.
- не проводилось исследований по анализу ошибок нейронного МП с русского языка, связанных со спецификой той или иной предметной области или с особенностями языка.

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.

Качественный перевод — передача смысла текста на языке оригинала в соответствии с языковыми и культурными традициями языка перевода согласно требованиями конечного реципиента.

Процесс перевода формально может быть представлен совокупностью множеств и операций над этими множествами. В дальнейшем слово «множество» опускается для удобства повествования. Пусть txt_{iTXT} – исходный текст, характеризующийся множеством выраженных смысловых единиц $СМ_i$, $яз_{вх}$ – язык исходного текста, $яз_{вых}$ – язык перевода, $дп_{iДП}$ – предметная область исходного текста, $TP|txt_{iTXT} : \exists ОК|TP \in R$ – требования к переводу исходного текста на язык перевода, такие, что существует $ОК$ – аналитически вычислимая оценка качества перевода, соответствующая критериям, зафиксированным в требованиях к переводу, перевод выполняется переводчиком $пер_{iПЕР}$, обладающим компетенцией $\overrightarrow{Кпер_{iПЕР}}$ и специализацией $\overrightarrow{Спер_{iПЕР,}дп_{iДП}}$, тогда качественный перевод – это сгенерированный текст txt_{jTXT} на языке $яз_{вых}$, характеризующийся множеством выраженных смысловых единиц $СМ_j$: $(СМ_j \rightarrow СМ_i)$. В общем виде соотношения множеств функция перевода $F_{пер}$ может быть описана как

$$F_{пер} : (txt'_{iTXT}, \overrightarrow{КперiПЕР}, \overrightarrow{СперiПЕР,дniДП}) \rightarrow txt_{jTXT} \mid (СМj \rightarrow СМi,\ TP|txt_{iTXT}). \quad (1)$$

Обобщенно модель процесса перевода представлена на рисунке 1.

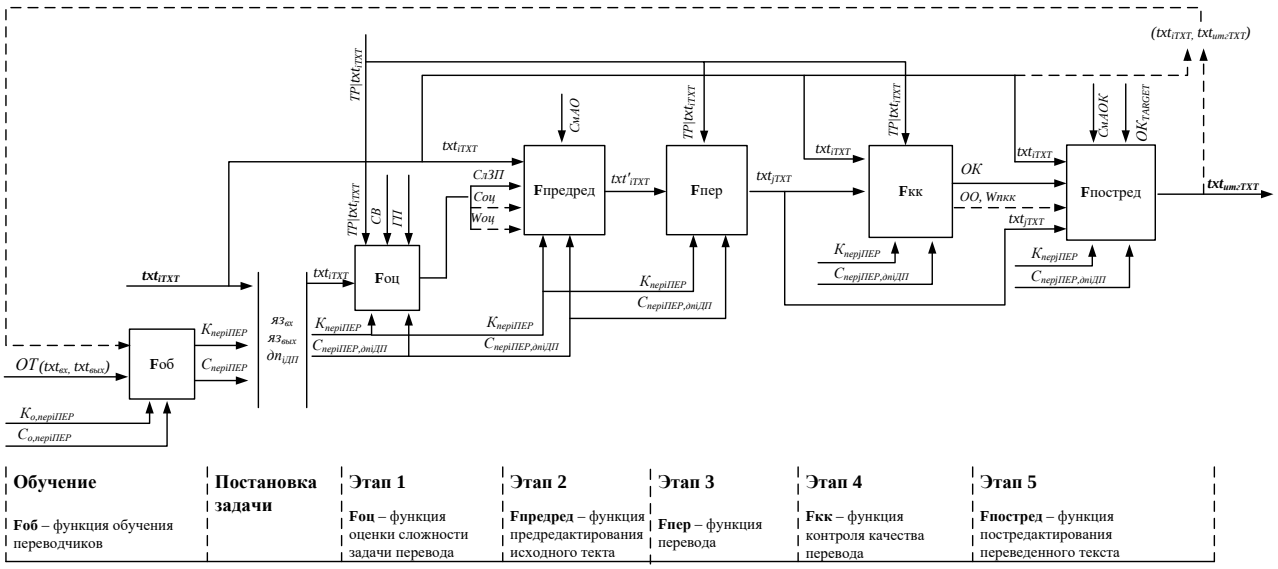

Рисунок 1 – Модель процесса перевода

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 OK_{itrg}, определяющее принадлежность элементов множества OK_{itrg} классу «Высокая оценка качества перевода текста txt_{iTXT}»*:*

$$КП_i = \{(OK, \mu_{КПi}(OK))|OK \in OK_{itrg}\};$$

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

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

$$\mu_{КПi}(OK) = \frac{1}{1+e^{-\left(\frac{OK-OK_{доп}}{maxOK-OK_{доп}}\right)2\pi}} \qquad (2)$$

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

$$F_{МП}(\underline{\theta}, \mu_{КПi}(OK)) = lnP_{\underline{\theta}}(\underline{max}\ \mu_{КПi}(OK)) \rightarrow \max_{\theta}, \qquad (3)$$

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

Решение поставленной задачи лежит в области оптимизации и совершенствования алгоритмов генерации переведенного текста.

В результате теоретического моделирования определено, что 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 оптимизации предредактирования исходного текста с целью повышения вероятности получения качественного перевода.

Для разработки стратегии и методики оптимизационного предредактирования требуется определить критерий оптимизации исходного текста. В качестве такого критерия была выбрана оценка сложности 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

$\underline{СВ|txt_{iТХТ}}$ и $\underline{ГП|txt_{iТХТ,}}$ 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 $\underline{ТР|txt_{iТХТ}}$.

Оценки сложности задачи перевода включает следующие шаги:

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 $\underline{ТР|txt_{iТХТ}}$, 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 $\overline{W_{оц}}$ 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 $\overline{W_{оц}}$ формируется уравнение поиска теоретического результирующего фактора, т.е. качества перевода $\widehat{КП}$:

$$\widehat{КП}_i = w_0 + w_{1оц}C_{оцi_1} + w_{2оц}C_{оцi_2} + \cdots + w_{оцk}C_{оцik} \qquad (4)$$

Для системы МП веса значимости оценок рассчитываются на основании тренировочных данных с использованием численного метода наименьших квадратов, при котором минимизируется сумма квадратов отклонений эмпирических значений результативного признака от теоретических, полученных по уравнению (4):

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

$$S(w) = \sum_{i=1}^{R} (ОК_i - w_0 + w_{1оц}C_{оцi_1} + w_{2оц}C_{оцi_2} + \cdots + w_{оцk}C_{оцik})^2 \to min \qquad (5)$$

где R – объем тренировочной выборки.

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

$$\sum_{i=1}^{R} (ОК_i - \widehat{ОК}_i(C_{оцi}, w)) \frac{\partial \widehat{ОК}_r(C_{оцi}, w)}{\partial w} = 0 \qquad (6)$$

Получаем систему нормальных уравнений с k неизвестными:

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

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

$$W_{оц}=(C_{оц}{}^{T}\cdot C_{оц})^{-1}\cdot C_{оц}{}^{T}\cdot OK = (\frac{1}{R}C_{оц}{}^{T}\cdot C_{оц})^{-1}\frac{1}{R}C_{оц}{}^{T}\cdot OK \quad (7)$$

Шаг 5. Для каждого *i-го* фрагмента текста рассчитывается вероятность получения переведенного текста на таком уровне качества, который определяется требованиями $TP|txt_{iTXT}$, применив к уравнению (6) логит-преобразование:

$$p_i = \frac{1}{1+e^{-\widehat{КП}_i}} \quad (8)$$

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

$$СлЗП_i = \frac{1}{p_i}. \quad (9)$$

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

$$СлЗП_{txt_{iTXT}} = max\, СлЗП_i \quad (10)$$

В зависимости от значения $СлЗП_{txt_{iTXT}}$ определяется стратегия дальнейшей обработки текста, в том числе необходимость применять оптимизационное предредактирование.

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.

Опишем условия задачи, 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 $ТХТ_{isrc}$ и $мСлЗП_{isrc}$ биективны: $ТХТ_{isrc} \leftrightarrow мСлЗП_{isrc}$;

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 $ТХТ_{isrc}$ 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):

$$\mu_{нСлЗПi}(СлЗП) = \frac{1}{1+e^{(\frac{СлЗП-СлЗП_{доп}}{|minСлЗП-СлЗП_{доп}|})2\pi}} \quad (11)$$

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 $\mu_{нСлЗПi}(СлЗП) = \{\mu_0, \mu_1, \ldots, \mu_l\}$ в рассматриваемой серии экспериментов равняться $\{max\ \mu_{нСлЗПi}(СлЗП)_0, max\ \mu_{нСлЗПi}(СлЗП)_1, \ldots, max\ \mu_{нСлЗПi}(СлЗП)_l\}$. Эта вероятность меняется в зависимости от Ψ:

$$F_{АОПР}(\Psi, \mu_{нСлЗПi}(СлЗП)) = \prod_{l=1}^{L} F_{АОПР}(\mu_{нСлЗПi}(СлЗП)_l) =$$
$$P_\Psi(\mu_0 = max\ \mu_{нСлЗПi}(СлЗП)_0)\cdot\ldots\cdot P_\Psi(\mu_l = max\ \mu_{нСлЗПi}(СлЗП)_l) = \quad (12)$$
$$P_\Psi(\mu_0 = max\ \mu_{нСлЗПi}(СлЗП)_0),\ldots, \mu_l = max\ \mu_{нСлЗПi}(СлЗП)_l),$$

где l – номер экземпляра в обучающей выборке объемом L.

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

$$L_{АОПР}(\Psi, \mu_{нСлЗПi}(СлЗП)) = lnP_\Psi(max\ \mu_{нСлЗПi}(СлЗП)), \quad (13)$$

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:

$$\widehat{\Psi} = arg\ \max_{\Psi} L_{АОПР}(\Psi, \mu_{нСлЗПi}(СлЗП)) \quad (14)$$

Решение поставленной задачи оптимизации выполняется методом градиентного спуска (подъема).

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:

$$\nabla L_{АОПР}(\Psi) = (\partial L_{АОПР}/\partial\psi_1, \partial L_{АОПР}/\partial\psi_2, \ldots, \partial L_{АОПР}/\partial\psi_m), \quad (15)$$

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

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

$$\Psi^{[s+1]} = \Psi^{[s]} + \alpha \cdot \nabla L_{АОПР}(\Psi^{[s]}), \quad (16)$$

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 $\nabla L_{АОПР}(\Psi)$:

$$\| \nabla L_{АОПР}(\Psi) \| = \sqrt{\left(\frac{\partial L_{АОПР}}{\partial \psi_1}\right)^2 + \left(\frac{\partial L_{АОПР}}{\partial \psi_2}\right)^2 + \ldots + \left(\frac{\partial L_{АОПР}}{\partial \psi_m}\right)^2}. \quad (17)$$

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:

$$\left\| \nabla L_{АОПР}(\Psi^{[s]}) \right\| \leq \varepsilon. \quad (18)$$

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.

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 пере-

вода с точки зрения теории переводоведения, можно сказать, что это ряд преобразований (переводческих трансформаций), с помощью которых осуществляется перестроение от единиц оригинала к единицам перевода. 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 настройки модели оптимизационного предредактирования:

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 тренировка НМП

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:

$$TrainCor = \{(src\_ref_i; pre\_src_i): \exists (tgt1_i, tgt2_i) \mid QC\_score(tgt2_i) > QC\_score(tgt1_i) \;\&\; \Delta QC\_score_i \geq d_{max}\} \quad (19)$$

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*.**

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

Для **расчёта 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 $С_{оцi}$ 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.), $ОП \subset СВ$; *лексические* (процент покрытия текста лексическими минимумами, частотными списками, специфичность лексики и др.), $ЛП \subset СВ$; *морфологические* (лексические и грамматические свойства формы слова), $МП \subset СВ$; *синтаксические* (глубина глагольных и именных групп, связи между глаголами в предложениях), $СП \subset СВ$; *признаки, основанные на базовых подсчетах* (средняя длина слов и предложений и пр.), $БП \subset СВ$.

$$СВ = ОП \cup ЛП \cup МП \cup СП \cup БП \quad (20)$$

Для морфологического и синтаксического разбора текста используем схему Universal Dependencies, 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

$$\forall\, UD_n \;\exists\; C_{оцi,k} = \frac{\sum_l^m 1|\, f(t_l)=UD_n}{m} \quad (21)$$

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 $C_{оцi}$. 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 6, 7, 8 рассчитывается сложность задачи перевода *СлЗП* для каждого *i-го* фрагмента текста.

Затем происходит отбор фрагментов текста, которым соответствуют высокие значения $СлЗП_i$, то есть

$$txt_i^* \in txt_{iTXT} \mid СлЗП_i > СлЗП_{доп}, \quad (22)$$

где $СлЗП_{доп}$ – допустимое значение сложности задачи перевода.

Далее производим редактирование текста в соответствии с имеющимися методами и алгоритмами оптимизационного предредактирования по критерию минимизации *СлЗП*.

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 машинного перевода текстов с русского языка на английский язык; препроцессинга текстовых данных для взвешенной оценки параметров русскоязычного текста; вероятностной оценки сложности переводческой задачи для систем машинного перевода; 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 машинного перевода. 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.

Архитектура программного комплекса представлена на рисунке 2.

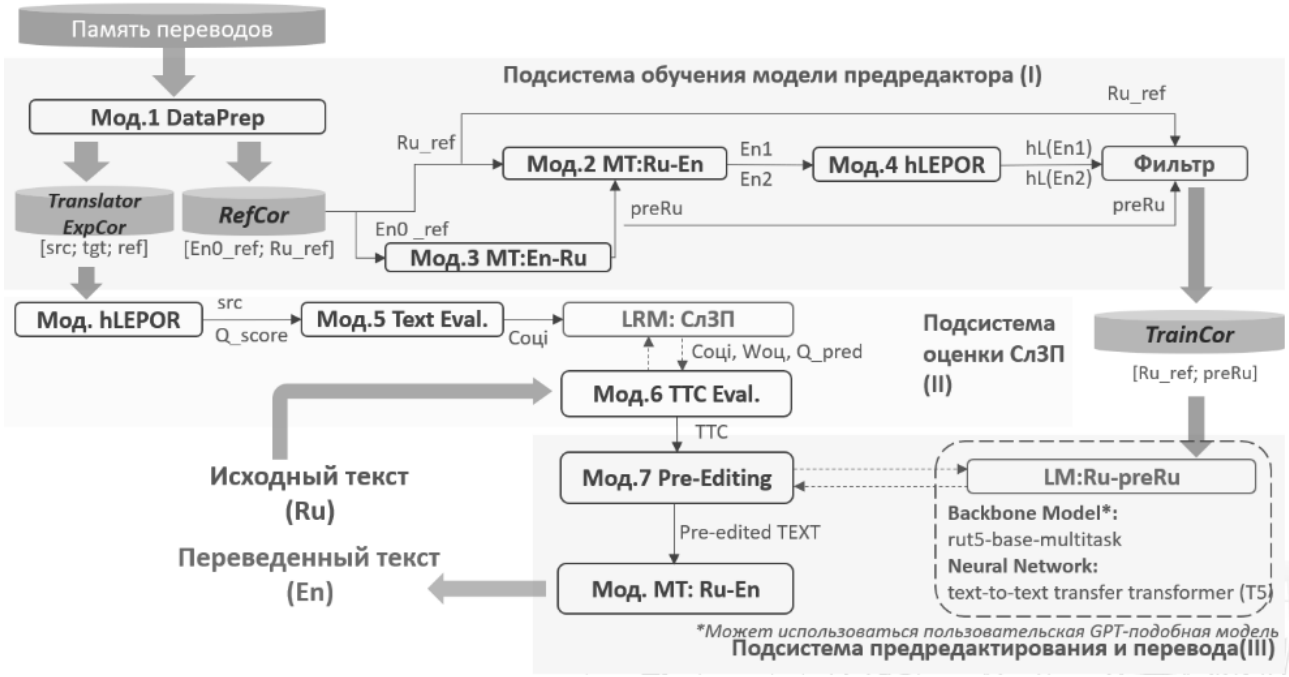

Рисунок 2 – Архитектура программного комплекса для повышения качества МП русскоязычных текстов на английский язык путем оптимизационного предредактирования

Структура исходных корпусов, используемых для настройки и тестирования языковой модели *LM:Ru-preRU*:

RefCor: [текст на английском языке (оригинал в базах памяти переводов) *En0_ref*; текст на русском языке (перевод, выполненный носителем) *Ru_ref*]. Объем корпуса – 139 438 пар предложений.

TestCor: [текст на русском языке (оригинал в базах памяти переводов) *Ru_test*; текст на английском языке (ручной перевод, проверенный редактором) *En0_test*]. Объем корпуса – 16 707 пар предложений.

В результате обработки эталонного корпуса *RefCor* модулями системы 1-4 был получен тренировочный корпус *TrainCor* объемом 88 631 строк. Для реализации модуля 7 использована базовая модель русского языка rut5-base-multitask на основе T5. Эта модель предобучена на широком спектре задач обработки и генерирования русскоязычных текстов и была дообучена на корпусе *TrainCor*, а затем интегрирована в систему оптимизационного предредактирования. Модуль 4 реализует алгоритм оценки качества машинного перевода по метрике hLEPOR, который оценивает «близость» выполненного перевода к эталонному тексту.

Схематичное представление методики тестирования языковой модели и модуля оптимизационного редактирования представлено на рисунке 3.

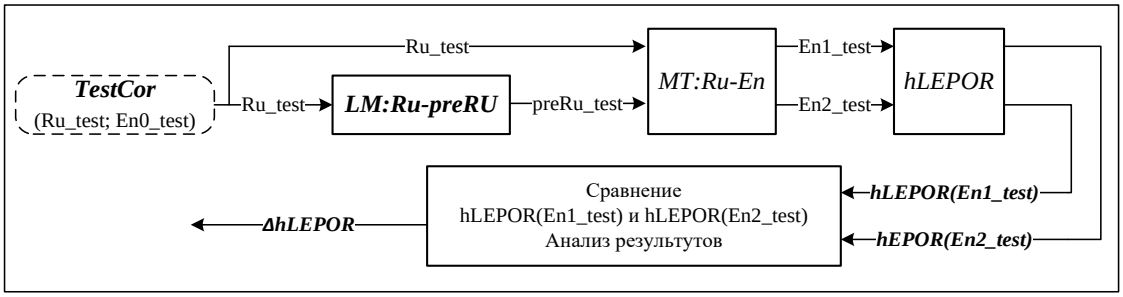

Рисунок 3 – Схематичное представление методики тестирования языковой модели и модуля оптимизационного редактирования

При реализации описанных методик оценки сложности задачи перевода и тренировки модели оптимизационного предредактирования приняты следующие допущения:

1) Критерий качества перевода должен быть четко определен и формализован с возможностью получения вещественного значения. Могут применяться любые метрики оценки качества в зависимости от требований к качеству перевода. В рамках исследования была выбрана метрика hLEPOR, которая имеет наивысший балл корреляции Пирсона с человеческими суждениями по языковой паре английский-русский.

2) Для тестирования переводчика 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 переводчиком/системой МП и проверенный перевод, утвержденный редактором, происходит автоматически в режиме реального времени.

В целях реализации представленной методики оценки сложности задачи перевода разработан препроцессор для взвешенной оценки свойств русскоязычного текста, включая морфологические, синтаксические, лексические и др., всего 96 параметров.

Тестирование программного комплекса проводилось на корпусе *TestCor*. Показано, что нецелесообразно применять оптимизационное редактирование ко всем текстам без использования критерия сложности задачи перевода, так как при этом среднее качество перевода значительно снижается. С использованием оценки сложности задачи перевода было отобрано 4071 экземпляр для оптимизационного предредактирования (24,4% тестовой выборки), для которых удалось добиться повышения качества машинного перевода на 15-30% по показателю hLEPOR.

Программный комплекса внедрен в контур автоматизации процессов перевода ООО «Агентство переводов «ФИАС-Амур» (г. Комсомольск-на-Амуре). 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%.

Пример предредактирования и его влияния на значения параметров текста и качество перевода на английский язык представлены в таблицах 1,2. В таблицах представлены наиболее значимые параметры для соответствующего текста.

Таблица 1 – Пример предредактирования русскоязычного текста – а) исходный текст; б) текст после предредактирования

Текст 1	**hLEPOR**	**ADP**	**conj**	**punct**
а) В результате многодневной переписки между представителями арендатора, арендодателя, сервисной компании и завода производителя результат о ремонте или замене станции достигнут не был.	0,5375	0,1200	0,1600	0,1200
б) В результате переписки между представителями Арендатора, Арендодателя, Сервисной компании и Производителя, решение о ремонте или замене станции не было достигнуто.	0,7560	0,1250	0,1667	0,1667

Таблица 2 – Пример предредактирования русскоязычного текста – а) исходный текст; б) текст после предредактирования

Текст 2	hLEPOR	ADP	VERB	xcomp
а) Оборудование должно быть рассчитано на двойные фидеры, а если такое оборудование отсутствует, в центральном шкафу предусматривают установку контроллера автоматического ввода резерва.	0,5051	0,0833	0,1250	0,0417
б) Оборудование должно быть способно управлять двумя фидерами, в случае отсутствия такого оборудования в центральном шкафу должен быть установлен переключатель ввода резерва.	0,6512	0,0909	0,0909	0,1364

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.

ОСНОВНЫЕ РЕЗУЛЬТАТЫ РАБОТЫ

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.

ПУБЛИКАЦИИ ПО ТЕМЕ ДИССЕРТАЦИИ

Публикации в изданиях, рекомендованных ВАК России:

1 **Животова А.А.** Оптимизационное предредактирование узкоспециальных русскоязычных текстов для их машинного перевода на английский язык / А.А. Животова, В. Д. Бердоносов // Информационные и математические технологии в науке и управлении. – 2024. – № 2 (34). – С. 169-182.

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

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

Публикации в изданиях, индексируемых в базе Scopus:

4 **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.

5 **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.

6 **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.

Публикации в других изданиях:

7 **Животова, А. А.** Автоматизация предредактирования русскоязычных текстов с целью повышения качества их машинного перевода на английский язык / А. А. Животова, В. Д. Бердоносов // 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.

8 **Животова, А. А.** Стратегия предредактирования исходного текста на основании автоматической оценки сложности задачи перевода для повышения качества машинного перевода узкоспециальных текстов на английский язык / А. А. Животова, В. Д. Бердоносов // Компьютерная лингвистика и интеллектуальные технологии: по материалам ежегодной международной конференции «Диалог». – № 22. – Доп. том. – 2023. – С. 1141-1149.

9 **Животова, А. А.** Машинный перевод корпусов текста для прикладных и исследовательских задач / А. А. Животова, В. Д. Бердоносов // 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.

10 **Животова, А. А.** Автоматизированная оценка параметров русскоязычного текста / А. А. Животова, В. Д. Бердоносов // 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.

11 **Животова, А. А.** Практическое применение вероятностной оценки сложности задачи перевода / А. А. Животова, В. Д. Бердоносов // 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.

12 **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 конференции 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.

Свидетельства о государственной регистрации программ для ЭВМ и БД:

13 Свидетельство о государственной регистрации программы для ЭВМ № 2023682260. Программный комплекс для предредактирования и машинного пе-

ревода узкоспециальных русскоязычных текстов на английский язык / **А. А. Животова** (RU), В. Д. Бердоносов (RU) // Правообладатель: **А. А. Животова** (RU); заявл. 09.10.2023; зарегистр. 24.10.2023

14 Свидетельство о государственной регистрации программы для ЭВМ № 2023669254. Программа для тренировки языковой модели в решении задач перефразирования русскоязычных технических текстов / **А. А. Животова** (RU) // Правообладатель: **А. А. Животова** (RU); заявл. 25.08.2023; зарегистр. 12.09.2023

15 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 / **А. А. Животова** (RU) // Правообладатель: **А. А. Животова** (RU); заявл. 25.08.2023; зарегистр. 06.09.2023

16 Свидетельство о государственной регистрации программы для ЭВМ № 2023668511. Предредактор русскоязычных узкоспециальных текстов для систем машинного перевода / А**. А. Животова** (RU) // Правообладатель: **А. А. Животова** (RU); заявл. 25.08.2023; зарегистр. 29.08.2023

17 Свидетельство о государственной регистрации базы данных № 2023623022. База данных показателей структурного анализа предложений технических русскоязычных текстов / **А. А. Животова** (RU) // Правообладатель: **А. А. Животова** (RU); заявл. 25.08.2023; зарегистр. 1.09.2023

18 Свидетельство о государственной регистрации программы для ЭВМ № 2023665348. Программный модуль автоматической очистки сырых данных из памятей переводов CAT для тренировки моделей нейронного машинного перевода / **А. А. Животова** (RU) // Правообладатель: **А. А. Животова** (RU); заявл. 10.07.2023; зарегистр. 14.07.2023

19 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 / **А. А. Животова** (RU) // Правообладатель: **А. А. Животова** (RU); заявл. 27.06.2023; зарегистр. 28.06.2023

20 Свидетельство о государственной регистрации программы для ЭВМ № 2023617748. Программа для вероятностной оценки сложности переводческой задачи для систем машинного перевода / **А. А. Животова** (RU), В. Д. Бердоносов (RU) // Правообладатель: ФГБОУ ВО «КнАГТУ»; заявл. 28.02.2023; зарегистр. 13.04.2023.

21 Свидетельство о государственной регистрации программы для ЭВМ № 2023614410. Препроцессинг текстовых данных для взвешенной оценки параметров русскоязычного текста / **А. А. Животова** (RU) // Правообладатель: **А. А. Животова** (RU); заявл. 13.02.2023; зарегистр. 01.03.2023

22 Свидетельство о государственной регистрации программы для ЭВМ № 2023613906. Парсер машинного перевода узкоспециальных технических текстов с русского языка на английский язык / **А. А. Животова** (RU) // Правообладатель: **А. А. Животова** (RU); заявл. 13.02.2023; зарегистр. 21.02.2023.