Компьютерная лингвистика и интеллектуальные технологии:
по материалам международной конференции «Диалог 2023»

Москва, 14–17 июня 2023 г.

Pre-editing Strategy Based on Automatic Evaluation of Translation Complexity to Improve the Quality of Specialized Texts Machine Translation into English

Alena A. Zhivotova
Komsomolsk-na-Amure State University, Komsomolsk-na-Amure, Russia
zhivotova.aa@gmail.com

Victor D. Berdonosov
Komsomolsk-na-Amure State University, Komsomolsk-na-Amure, Russia
berd1946@gmail.com

Abstract

The study addresses the issue of applying optimizing pre-editing of Russian-language texts in order to improve the quality of machine translation into English. A probabilistic assessment of translation task complexity is proposed to be used for selecting a pre-editing strategy. A generalized model of the translation process is presented. A mathematical model and algorithm for automated assessment of translation task complexity are proposed. Test of the model on specialized texts of oil and gas industry is described, which showed that the estimate correlates with an estimate of translation quality and can be used in selecting a strategy for optimizing pre-editing of source texts in machine translation tasks.

Keywords: machine translation, optimizing pre-editing, translation task complexity, translation quality
DOI: 10.28995/2075-7182-2023-22-1141-1149

Стратегия предредактирования исходного текста на основании автоматической оценки сложности задачи перевода для повышения качества машинного перевода узкоспециальных текстов на английский язык

Алена Животова
Комсомольский-на-Амуре государственный университет, Комсомольск-на-Амуре, Россия
zhivotova.aa@gmail.com

Виктор Бердоносов
Комсомольский-на-Амуре государственный университет, Комсомольск-на-Амуре, Россия
berd1946@gmail.com

Аннотация

Исследование рассматривает вопрос применения оптимизационного предредактирования русскоязычных текстов с целью повышения качества машинного перевода на английский язык. Для выбора стратегии предредактирования предлагается использовать вероятностную оценку сложности задачи перевода. Представлена обобщенная модель процесса перевода. Предложены математическая модель и алгоритм автоматизированной оценки сложности задачи перевода. Описано тестирование модели на узкоспециальных текстах нефтегазовой тематики, которое показало, что данная оценка коррелирует с оценкой качества перевода и может быть использована при выборе стратегии оптимизационного предредактирования исходных текстов в задачах машинного перевода.

Ключевые слова: машинный перевод, оптимизационное предредактирование, сложность задачи перевода, качество перевода

1 Введение

Рассматривая вопрос качества машинного перевода (МП) для конечного реципиента следует учитывать следующие факторы: специфика предметной области и компетенция пользователя МП.

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 двуязычных корпусов проблематично ввиду ограничений конфиденциальности данных и секретности разработок. 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 критическое значение для коммуникации и обмена технологиями.

Поскольку для корректного использования систем МП необходимо прямое участие человека на всех этапах работы системы, справедливо, что для достижения оптимального результата работы, пользователь 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. Предоставляя пользователю средства обработки текста на языке, носителем которого он является, на любом из этапов перевода, можно повысить качество перевода. Зная ключевые параметры текста и их связь с предполагаемой оценкой качества, становится возможным предложить алгоритмы и инструменты редактирования текста с целью его оптимизации по критерию максимизации качества перевода.

Идея автоматического и полуавтоматического оптимизационного редактирования текста в задачах МП лежит в основе интерактивного МП. Основная масса работ и исследований в области интерактивного МП посвящены постредактированию, в том числе, его автоматизации [1, 2]. Предредактированию посвящено меньше работ, однако существующие статьи указывают на эффективность такого подхода к повышению качества перевода.

Так, Hiraoka и Yamada [3] в своей работе предприняли попытку сформулировать основные правила предредактирования текста при переводе с японского языка на английский. 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 [4], с французского на английский [5], с индонезийского на английский [6], с английского на испанский [7]. Все указанные выше исследования показали, что предредактирование и переписывание исходного текста, опираясь на правила, повышает качество МП. Seretan, Bouillon и Gerlach в своем исследовании [8] показали, что использование даже простых полуавтоматических правил предредактирования текста повышает качество статистического МП. Исследование Шей [9] в языковой паре китайский-английский показало преимущества и позитивное влияние предредактирования на качество МП для пользователей с низким уровнем владения языком перевода в сравнении с постредактированием. Отмечается необходимость анализа исходного текста с целью определить «уязвимости» с точки зрения применяемой технологии МП, на основе которого становится возможной разработка правил предредактирования. Gerlach, Porro, Bouillon и Lehmann [5] показали, что такой подход позволяет сократить время пост-редактуры МП в два раза и в 65% случаев повышает качество результата МП.

Цель исследования – разработать стратегию оптимизационного предредактирования исходных текстов для повышения качества машинного перевода узкоспециальных текстов с русского на английский язык. Решение задачи повышения качества перевода узкоспециальных текстов позволит оптимизировать затраты на перевод, повысить надежность существующих систем, снизить зависимость качества перевода от человеческого фактора.

2 Модель процесса перевода

Для решения поставленной задачи проведено подробное математическое моделирование процесса и основных понятий перевода на основе теории множеств и построение процессной модели перевода. Обобщенная модель процесса перевода представлена на рисунке 1.

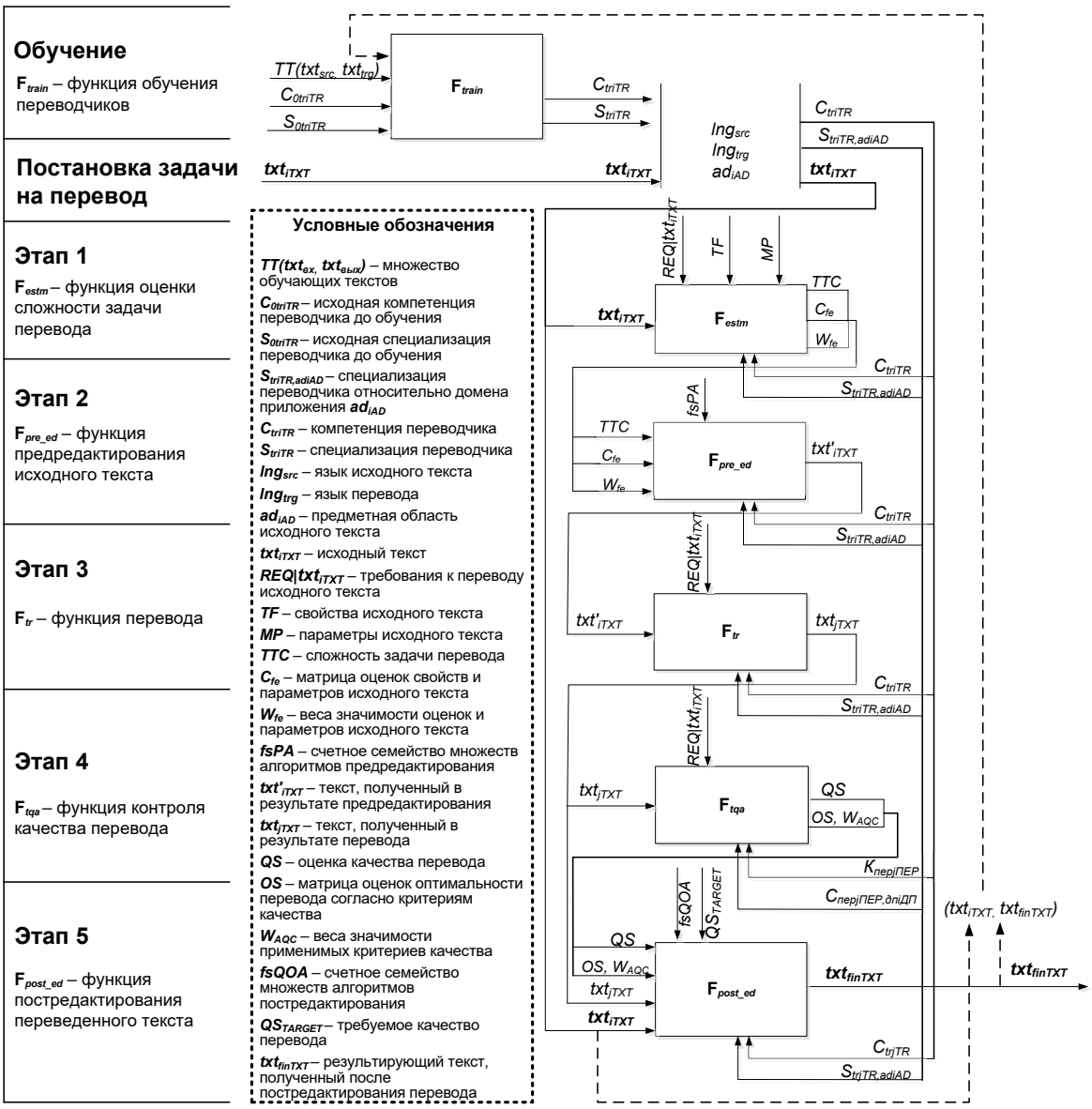

Рисунок 1: Обобщенная модель процесса перевода

В результате 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 исходного текста с целью повышения вероятности получения качественного перевода. Например, он расшифровывает аббревиатуры, применяет переводческие трансформации к исходному тексту и т.д. Результат работы МП – черновик, который пользователь должен оценить и доработать самостоятельно. При этом он должен обладать высоким уровнем компетенций в языке перевода, что ограничивает применение МП, особенно в узкоспециальных областях знаний, для которых не собраны достаточные корпуса для тренировки моделей перевода, и для которых требуется более тщательная проверка переведенного текста. Для разработки стратегии оптимизационного предредактирования была поставлена задача в первую очередь разработать способ оценки сложности задачи перевода и предполагаемого качества перевода для систем МП.

3 Вероятностная оценка сложности задачи перевода

Переводчик/система МП tr_{iTR} получает текст txt_{iTXT} на языке lng_{src} для перевода на язык lng_{trg} и требования к переводу $REQ|txt_{iTXT}$. При оценке сложности задачи перевода переводчик обращает внимание на неизвестные ему слова и сочетания слов на языке lng_{src}, для которых он не может идентифицировать значение смысловой единицы, либо смысловые единицы, для которых он не может найти аналог на языке перевода lng_{trg} среди известных ему слов и сочетаний слов.

Множества свойств и параметров исходного текста $TF|txt_{iTXT}$ и $MP|txt_{iTXT,}$ и то, обладает ли переводчик достаточной компетенцией $\overrightarrow{Ctr_{iTR}}$ относительно языков lng_{src} и lng_{trg} и специализацией $\overrightarrow{Str_{iTR},ad_{iAD}}$, т.е. навыками описания семантических единиц на языке перевода в рамках заданной предметной области исходного текста ad_{iAD}, определяет вероятность создания переводчиком переведенного текста на таком уровне качества, который определяется требованиями $REQ|txt_{iTXT}$.

Алгоритм оценки сложности задачи перевода:

Шаг 1. Исходя из домена приложения текста ad_{iAD}, формируется множество оценок текста $FE = TF \cup MP$.

Шаг 2. Для каждого значения $tf_{iTF}, mp_{iMP} \in FE$, на основе требований к переводу $REQ|txt_{iTXT}$, компетенций переводчика относительно языковой пары $\overrightarrow{Ctr_{iTR}}$ и специализации переводчика относительно домена приложения текста $\overrightarrow{Str_{iTR},ad_{iAD}}$ формируется значение значимости w_{fek}, множество нормированных значений w_{ouk} значимости формируют матрицу значимости оценок сложности $\overline{W_{fe}}$ 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 C_{fei} размерностью $1 \times k$, где k – общее число оценок.

Шаг 4. На основании оценок C_{fei} и значимости $\overline{W_{fe}}$ формируется уравнение поиска теоретического результирующего фактора, т.е. качества перевода $\widehat{TQ}$:

$$\widehat{TQ}_i = w_0 + w_{1fe}C_{fei_1} + w_{2fe}C_{fei_2} + \cdots + w_{fek}C_{feik} \quad (1)$$

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 (1) логит-преобразование [11]:

$$p_i = \frac{1}{1+e^{-\widehat{TQ}_i}} \quad (2)$$

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

$$TTC_i = \frac{1}{p_i}, \quad (3)$$

где p_i – это вероятность создания переводчиком перевода требуемого качества, рассчитанная по формуле (2).

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

$$TTC_{txtiTXT} = max\ TTC_i \quad (4)$$

Алгоритм реализации описанной математической модели в рамках прикладной задачи перевода:

1. Тестирование переводчика/системы МП на тренировочном корпусе текстов заданной тематики, для которых имеется эталонный перевод.
2. Вещественная оценка параметров текста в тренировочном корпусе.
3. Оценка качества выполненного переводчиком тестового перевода.
4. Поиск весов значимости параметров текста заданной тематики для тестируемого переводчика.
5. Получение уравнения поиска теоретического качества перевода, выполненного тестируемым переводчиком. На основе уравнения производятся расчеты ожидаемого качества, вероятность получения перевода, соответствующего классу «качественный перевод» и сложность задачи перевода.

Данный алгоритм протестирован на реальных данных в рамках перевода текстов нефтегазовой тематики системой МП.

4 Реализация модели оценки сложности задачи перевода

4.1 Тестирование переводчика

Для тестирования разработанного алгоритма был разработан программный модуль-парсер с использованием модели нейронного МП на базе Transformers и Open Source модели Helsinki-NLP[1] для языковой пары русский-английский, предварительно обученной на корпусе OPUS[2]. Разработка и инициализация парсера МП производилась в среде Google Colab. Использование парсера обусловлено необходимостью протестировать алгоритм на большом массиве данных, не прибегая к использованию коммерческих программ ввиду ограничений на конфиденциальность используемых исходных данных.

Тренировочный корпус – корпус двуязычных текстов, собранный на базе Translation Memories из сред автоматизации перевода, применяемых компанией-поставщиком лингвистических услуг, специализирующейся на переводе технических текстов в сфере нефтегазопереработки. Структура данных: [исходный текст, ручной перевод, проверенный редактором]. Корпус предварительно очищен от шумов, таких как сегменты, содержащие нетекстовую информацию, теги форматирования текста, строки длиной менее 50 символов. Объем корпуса составил ~1 744 400 токенов или ~7 300 стандартных страниц[3] русскоязычного текста.

4.2 Вещественная оценка параметров исходного текста

Для вещественной оценки параметров текста был разработан программный модуль для оценки параметров русскоязычного текста по четырем группам признаков: морфологические, синтаксические, лексические и прочие признаки. Программа разбивает текст на токены и рассчитывает значение для 96 вещественных признаков.

Для оценки морфологических и синтаксических признаков текста используется разбор по схеме универсальных зависимостей [12]. 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. Оценивается доля слов исследуемой части текста по каждому тегу части речи, как наиболее информативный морфологический признак. Схема универсальных зависимостей позволяет производить синтаксический разбор текста, маркируя каждую его единицу (токен) соответствующим тегом отношения «rel=».

Морфологический и синтаксический разбор включает следующие шаги:

1. Разбиение текста на токены.
2. Определение для каждого токена значений свойств pos (часть речи) и rel (роль в предложении).
3. Подсчет количество частей речи в строке и число токенов по каждому тегу.
4. Расчет для всех тегов отношения количества токенов с соответствующим тегом к общему числу токенов в предложении.

Морфологический и синтаксический анализ выполняется при помощи библиотеки natasha[4] для моделирования НЛП на основе глубокого обучения для русского языка, который имеет сравнимую точность с большими моделями BERT SOTA, но занимают в 50 раз меньше места.

Для оценки лексических признаков использовался Национальный частотный словарь русской лексики [13], который был предварительно обработан и отсортирован по значению частотности слов. Оценка производилась для токенов, включающих слова, которые предварительно лемматизированы при помощи библиотеки natasha.

Оценка факторов на нормальность распределения показала наличие большого количества факторов с экспоненциальным распределением значений и факторов, большая часть наблюдений по которым равна 0. Будем учитывать это при оценке точности модели, так как нормализация и

[1] https://huggingface.co/Helsinki-NLP
[2] https://opus.nlpl.eu/
[3] Стандартная страница равна 1800 знаков с пробелами
[4] https://github.com/natasha

шкалирование данных не позволит устранить нулевые значения, а балансировка данных по нулевым значениям может привести к значительным потерям данных по другим факторам.

4.3 Критерий качества машинного перевода

В рамках представленного алгоритма могут применяться любые метрики оценки качества в зависимости от требований к качеству перевода. Была выбрана метрика hLEPOR, которая является комбинацией существующих и доработанных факторов и показывает лучшие результаты оценки по сравнению с MPF, ROSE, METEOR, BLEU и TER, а также имеет наивысший балл корреляции Пирсона с человеческими суждениями по языковой паре английский-русский [14]. Оценка метрики производилась путем сравнения сгенерированного МП (гипотезы) и эталонным переводом, выполненным человеком, при помощи библиотеки hLEPOR[5]. Диапазон изменения метрики от 0 до 1, где 0 – полное несовпадение гипотезы с эталоном, а 1 – полное совпадение.

4.4 Корреляционная связь качества перевода с параметрами исходного текста

Для поиска весов значимости параметров текста и получения многофакторного уравнения поиска теоретического качества перевода использовалась модель логистической регрессии, реализованная в библиотеках statsmodels.api и sklearn для языка Python. Результаты моделирования представлены на рисунке 2.

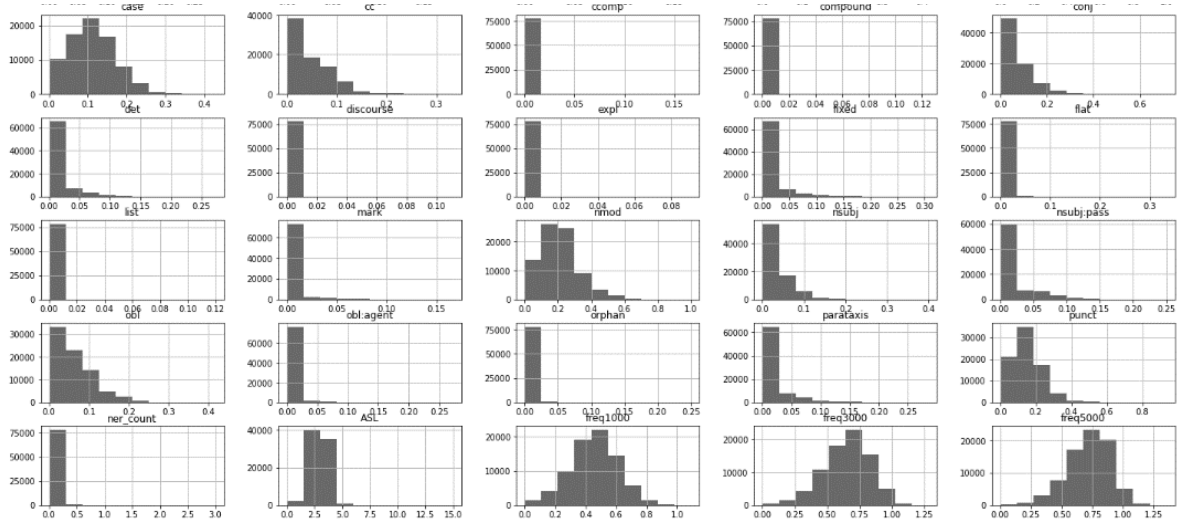

Рисунок 2: Гистограммы распределения данных для части факторов

[5] https://pypi.org/project/hLepor

	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 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. Всего выявлено 26 параметров, из которых 12 имеют отрицательную зависимость с потенциальной оценкой качества перевода и являются мешающими. 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 показателей ROC-AUC [15] и коэффициента корреляции Пирсона, рассчитанного для *TTC* и фактической *QS* для 85125 строк текста, согласно полученной зависимости (таблица 1).

ROC-AUC показывает, что на основании признаков исходного текста без оценки его семантики, возможно предсказывать ожидаемое качество МП и имеется потенциал доработки модели и повышения ее точности с учетом разреженности пространства факторов и распределения их значений. Коэффициент $r_{TTC,QS}$ указывает на обратную корреляционную связь с выбранной оценки качества перевода и сложности задачи перевода.

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.

Показатель	**Значение**
ROC-AUC	0,6256
$r_{TTC,QS}$	−0.26

Таблица 1: Результаты моделирования

5 Стратегия предредактирования исходных текстов

Предредактирование – процесс модификации проблемных фрагментов исходного текста за счет изменения его структуры, состава лексических единиц и т.п. при сохранении семантической эквивалентности с целью оптимизации сложности переводческой задачи с учетом влияния такой модификации на качество перевода.

Предредактирование выполняется при средней и высокой сложности переводческой задачи с использованием разнообразных алгоритмов, например, по критериям: минимизации $СлЗП_{txt_{iTXT}}$, минимаксимизации отдельных свойств и параметров исходного текста и др.

Диапазоны значений $TTC_{txt_{iTXT}}$, соответствующие низкому, среднему и высокому уровням определяются на основе предварительного анализа требований к переводу и с учетом способа перевода (ручной/машинный). Цель предредактирования – снизить сложность переводческой задачи $TTC_{txt_{iTXT}}$ до низкой.

На начальном этапе предредактирования происходит отбор фрагментов текста, которым соответствуют высокие значения TTC_i, то есть

$$txt_i^* \in txtiTXT \mid TTC_i > TTC_{alw} \quad (5)$$

где TTC_{alw} – 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_{fei}, значения которых повышают сложность переводческой задачи данного фрагмента: $C_{fei}|TTC_i > TTC_{alw}$.

Для каждого найденного элемента матрицы C_{fei}, соответствующего условию $C_{fei}|TTC_i > TTC_{alw}$., в зависимости от его значения и ограничений алгоритмов предредактирования определяется стратегия предредактирования (таблица 2). Выбор алгоритма зависит от критерия оптимизации выбранного значения.

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

Таблица 2: Стратегии предредактирования

Далее производится редактирование текста в соответствии с имеющимися методами и алгоритмами предредактирования, формируется текст txt'_{iTXT} и происходит переход к этапу перевода.

Предредактирование позволяет повысить качество МП. Пример предредактирования и его влияния на значения параметров текста и качество перевода на английский язык представлены в таблице 3. В таблице представлены наиболее значимые параметры для соответствующего текста.

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
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

Таблица 3: Пример предредактирования русскоязычного текста – а) исходный текст; б) текст после предредактирования

6 Заключение

В рамках исследования впервые разработана математическая модель и алгоритм автоматизированной вещественной оценки сложности задачи перевода и подтверждена обратная корреляционная связь такой оценки с оценкой качества, рассчитанной по методу hLEPOR для узкоспециальных текстов нефтегазовой тематики. Выявлены параметры русскоязычного узкоспециального текста, имеющие корреляционную связь с потенциальным качеством МП на английский язык по целевому показателю оценки метрики hLEPOR.

Подход может быть масштабирован на ручной перевод и внедрен в компаниях, генерирующих от 1000 страниц перевода в месяц, так как намечает подходы к управлению рисками, связанными с качеством перевода в зависимости от компетенции выбранных исполнителей, и предоставит индустрии инструмент объективной оценки исполнителей в рамках поставленной задачи на перевод.

На основе методики оценки сложности задачи перевода становится возможным автоматически определять стратегию предварительного оптимизационного редактирования текста с целью приведения значений его параметров к оптимальным, при которых вероятностная оценка качества МП стремится к максимальной. Дальнейшее исследование целесообразно направить на разработку методики расчета целевых диапазонов значений *TTC* и оптимальных значений параметров исходного текста по критерию максимизации качества перевода, а также разработку комплекса алгоритмов оптимизационного предредактирования.

References

[1] Yamada M. The impact of Google neural machine translation on post-editing by student translators // The Journal of Specialised Translation. — 2019. — vol. 31. — P. 87–106.
[2] Toledo Báez M. Machine translation and post-editing: impact of training and directionality on quality and productivity // Revista Tradumàtica. Tecnologies de la Traducció. — 2018. — vol. 16. — P. 24–34.
[3] Hiraoka Y.,Yamada M. Pre-editing plus neural machine translation for subtitling: effective pre-editing rules for subtitling of TED Talks // MT Summit XVII. — Dublin, Ireland. — 2019. — vol.2. — P. 64–74.
[4] 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), User studies papers. — Prague, Czech Republic. — 2017.
[5] Gerlach J., O'Brien S. et al. Combining pre-editing and post-editing to improve SMT of user-generated content // Proceedings of MT Summit XIV Workshop on Post-editing Technology and Practice. — Nice, France. — 2013. — P. 45-53.
[6] Taufik A. Pre-editing of Google neural machine translation // Journal of English Language and Culture. — 2020. — vol. 10. — No. 2. — P. 64–74.
[7] Mercader-Alarcón J., Sánchez-Martínez F. Analysis of translation errors and evaluation of pre-editing rules for the translation of English news texts into Spanish with lucy LT // Revista Tradumàtica. Tecnologies de la Traducció. — 2016. — vol. 14. — P. 172–186.
[8] Seretan V., Bouillon P. et al. A large-scale evaluation of pre-editing strategies for improving user-generated content translation // Proceedings of the Ninth International Conference on Language Resources and Evaluation. — Reykjavik, Iceland. — 2014. — P. 1793–1799.
[9] Shei Chi-Chiang. Teaching MT through pre-editing: three case studies // Proceedings of the 6th EAMT Workshop: Teaching Machine Translation. — Manchester, England. — 2002.
[10] 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». — 2021. — С. 191-204.
[11] Hosmer, D. W., Lemeshow, S. Applied Logistic Regression. 2nd edn. — New York: Wiley Chichester, 2000.
[12] Люкина Е. В. Использование универсальных зависимостей при грамматическом разборе многоязычного текста (на примере безличного предикатива) // Вестник Новосибирского государственного университета. Серия: Лингвистика и межкультурная коммуникация. — 2018. — Т. 16. — № 2. — С. 19-33.
[13] Ляшевская О. Н., Шаров С. А. Частотный словарь современного русского языка. — Москва : Азбуковник, 2009.
[14] Li-Feng Han A., Wong D. F. et al. Language-independent Model for Machine Translation Evaluation with Reinforced // Proceedings of the Machine Translation Summit XIV. — Nice, France. — 2013. — P. 215-222.
[15] Бенгфорт Б., Билбро Р., Охеда Т. Прикладной анализ текстовых данных на Python. Машинное обучение и создание приложений обработки естественного языка. — Санкт-Петербург: Питер, 2019.