УДК 004.9:519.768
DOI:10.25729/ESI.2024.34.2.016

Оптимизационное предредактирование узкоспециальных русскоязычных текстов для их машинного перевода на английский язык

Животова Алена Анатольевна, Бердоносов Виктор Дмитриевич

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

Аннотация. В статье исследуется процесс перевода текстов, а именно метод оптимизационного предредактирования, как способ повышения качества машинного перевода на английский язык при работе с русскоязычными узкоспециальными текстами. Авторы рассматривают математическую модель процесса перевода и постановку задачи машинного перевода, предлагают новую теорию вероятностной оценки сложности задачи перевода, приводят постановку и решение задачи оптимизационного предредактирования, описывают методику подготовки данных для обучения модели автоматического оптимизационного предредактирования. В результате исследования реализован программный комплекс оптимизационного предредактирования русскоязычных текстов. При разработке программного комплекса использованы ресурсы Центра коллективного пользования научным оборудованием «Центр обработки и хранения научных данных ДВО РАН». Данные для обучения и валидации моделей предоставлены ООО «Агентство переводов “ФИАС-Амур”». Тестирование программного комплекса показало эффективность предложенных методик для повышения качества машинного перевода узкоспециальных русскоязычных текстов на английский язык.

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

Цитирование: 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. – DOI:10.25729/ESI.2024.34.2.016.

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. Несмотря на выдающиеся прорывы нейросетевых, гибридных и больших языковых моделей машинного перевода (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 [1, 2]. 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. Именно на этот принцип опирается концепция интерактивного перевода, широко описанная в литературе [3]. Одним из направлений интерактивного перевода является перевод с предредактированием, когда исходный текст предварительно редактируется с целью его адаптации для более легкого «понимания» системой МП.

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 [4]; 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 пред-

редактированию [5, 6, 7]; L. O'Brien, D. Folaron, W. Aziz, M. Toledo – в области контролируемых и упрощенных языков [7]. Среди российских авторов и для русского языка данная тема освещена незначительно, однако известны работы И. В. Оборневой [8], А.Д Дмитриевой, А. Н. Лапошиной, М. Ю. Лебедевой [9] и др. в области оценки восприятия текста и упрощения русскоязычных текстов в соответствии с квалификацией реципиента.

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

Основная задача заключается в том, чтобы, используя особенности работы алгоритмов систем МП и основы теории перевода, автоматизировать предварительное редактирование исходных текстов, оптимизировав 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.

В основе проведенного исследования лежит применение апробированных математических методов, включая теорию множеств, численные методы оптимизации, такие, как метод наименьших квадратов и метод градиентного спуска, статистические методы, в том числе метод максимального правдоподобия.

В работе 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; 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; предложена новая архитектура и реализован программный комплекс для повышения качества МП 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 МП текстов с русского языка на английский язык.

1. Математическая модель процесса перевода. 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 $txt_{i\text{ТХТ}}$ – исходный текст, характеризующийся множеством выраженных смысловых единиц СМ_i, $\text{яз}_{\text{вх}}$ – язык исходного текста, $\text{яз}_{\text{вых}}$ – язык перевода, $\text{дп}_{i\text{ДП}}$ – предметная область (домен приложения) исходного текста, $\text{ТР}|txt_{i\text{ТХТ}} : \exists\ \text{ОК}|\text{ТР} \in R$ – 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 $\text{пер}_{i\text{ПЕР}}$, обладающим компетенцией $\mathbf{К}_{\mathbf{пер}i\mathbf{ПЕР}}$ и специализацией $\mathbf{С}_{\mathbf{пер}i\mathbf{ПЕР,\ дп}i\mathbf{ДП}}$, тогда качественный перевод – это сгенерированный текст $txt_{j\text{ТХТ}}$ на языке $\text{яз}_{\text{вых}}$, характеризующийся множеством выраженных смысловых единиц СМ_j: $(\text{СМ}_j \rightarrow \text{СМ}_i)$. В общем виде соотношения множеств функция перевода $F_{\text{пер}}$ может быть описана как

$$F_{\text{пер}} : (txt_{i\text{ТХТ}}, \mathbf{К}_{\mathbf{пер}i\mathbf{ПЕР}}, \mathbf{С}_{\mathbf{пер}i\mathbf{ПЕР,\ дп}i\mathbf{ДП}}) \rightarrow txt_{j\text{ТХТ}} \,|\big(\text{СМ}_j \rightarrow \text{СМ}_i, \text{ТР}\big|txt_{i\text{ТХТ}}\big). \quad (1)$$

Обобщенно модель процесса перевода представлена на рисунке 1. Подробное описание полной модели было опубликовано авторами ранее [10].

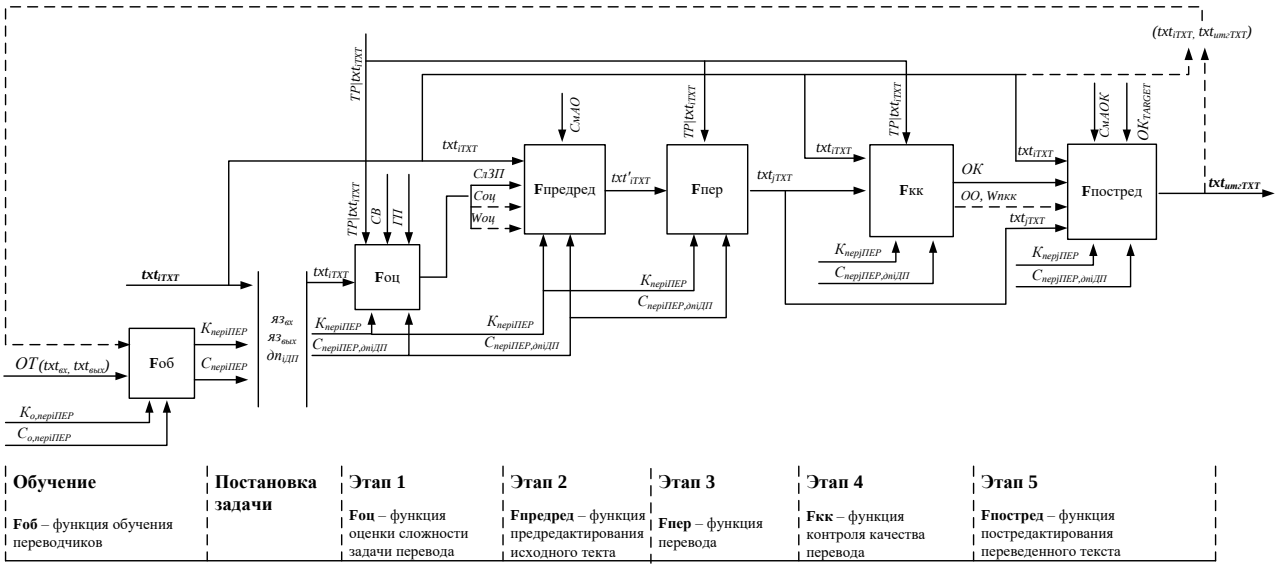

Рис. 1. Модель процесса перевода

Математически, для систем МП задачу перевода можно формализовать через метод максимизации функции правдоподобия [11] следующим образом. Опишем условия задачи, пусть:

$txt_{i\text{ТХТ}}$ – исходный текст на языке $\text{яз}_{\text{вх}}$.

$txt_{j\text{ТХТ}}$ – переведенный текст на языке $\text{яз}_{\text{вых}}$.

$TXT_{i\text{trg}}$ – множество всех возможных вариантов перевода текста txt_{iTXT} на язык $\text{яз}_{\text{вых}}$:

$$TXT_{i\text{trg}} = \{txt_0, txt_1, \dots, txt_j ,$$

где j – общее число вариантов перевода, $txt_{j\text{ТХТ}} \in TXT_{i\text{trg}}$.

$\text{ОК}_{i\text{trg}}$ – множество нормированных оценок качества перевода текста txt_{iTXT} в соответствии с требованиями к переводу $\text{ТР}|txt_{i\text{ТХТ}}$ для всех возможных вариантов перевода $TXT_{i\text{trg}}$:

$$\text{ОК}_{i\text{trg}} = \{\text{ОК}_0, \text{ОК}_1, \dots, \text{ОК}_j\},$$

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

Каждому варианту перевода соответствует одна оценка качества перевода, то есть множества $TXT_{i\text{trg}}$ и $\text{ОК}_{i\text{trg}}$ биективны: $TXT_{i\text{trg}} \leftrightarrow \text{ОК}_{i\text{trg}}$.

$[min\text{ОК};\ max\text{ОК}]$ – диапазон значений оценок качества перевода $\text{ОК}_{i\text{trg}}$.

$\text{ОК}_{\text{доп}}$ – минимально допустимое значение критерия «Высокая оценка качества перевода» при допущении, что чем выше значение ОК_j, тем лучше.

КП_i – нечёткое подмножество множества ОК_{itrg}, определяющее принадлежность элементов множества $\text{ОК}_{i\text{trg}}$ классу «Высокая оценка качества перевода текста $txt_{i\text{ТХТ}}$»:

$$\text{КП}_i = \{(\text{ОК},\ \mu_{\text{КП}i}(\text{ОК}))|\text{ОК} \in \text{ОК}_{i\text{trg}}\}.$$

$\mu_{\text{КП}i}(\text{ОК})$ – функция принадлежности, указывающая, в какой степени текст txt с оценкой ОК принадлежит нечеткому множеству КП_i .

$\mu_{\text{КП}i}(\text{ОК}) \in [0; 1]$ и имеет вид логистической кривой:

$$\mu_{\text{КП}i}(\text{ОК}) = \frac{1}{1+e^{-\left(\frac{\text{ОК}-\text{ОК}_{\text{доп}}}{max\text{ОК}-\text{ОК}_{\text{доп}}}\right)2\pi}}. \quad (2)$$

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

$$F_{\text{МП}}(\theta, \mu_{\text{КП}i}(\text{ОК})) = lnP_\theta(\max \mu_{\text{КП}i}(\text{ОК})) \to \max_\theta , \quad (3)$$

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

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.

2. Сложность задачи 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 $\text{яз}_{\text{вх}}$, для которых он не может идентифицировать значение смысловой единицы, либо смысловые единицы, для которых он не может найти аналог на языке перевода $\text{яз}_{\text{вых}}$ среди известных ему слов и сочетаний слов. Множества свойств и параметров исходного текста $\text{СВ}|txt_{i\text{ТХТ}}$ и $\text{ГП}|txt_{i\text{ТХТ}}$, и то, обладает ли переводчик достаточной компетенцией $\mathbf{К}_{\mathbf{пер}i\mathbf{ПЕР}}$ относительно языков $\text{яз}_{\text{вх}}$ и $\text{яз}_{\text{вых}}$ и специализацией $\mathbf{С}_{\mathbf{пер}i\mathbf{ПЕР},\ \mathbf{дп}i\mathbf{ДП}}$, 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 $\text{ТР}|txt_{i\text{ТХТ}}$.

Оценки сложности задачи перевода включает следующие шаги:

Шаг 1. Исходя из домена приложения текста $\text{дп}_{i\text{ДП}}$, формируется множество оценок текста $\text{ОЦ} = \text{СВ} \cup \text{ГП}$. Свойства текста СВ условно можно разделить на группы признаков: общие (количество символов/слов/строк и т.д., стиль, язык, домен приложения и пр.); лексические (процент покрытия текста лексическими минимумами, частотными списками и др.); морфологические (количество различных частей речи и грамматических форм); синтаксические (глубина глагольных и именных групп, связи между глаголами в предложениях); признаки, основанные на базовых подсчетах (средняя длина слов и предложений и пр.). Совокупность свойств и признаков определяет главные параметры текста ГП, к которым относится целостность, связность, удобочитаемость, сложность и другие.

Шаг 2. Для каждого значения $\text{св}_{i\text{СВ}}, \text{гп}_{i\text{ГП}} \in \text{ОЦ}$, на основе требований к переводу $\text{ТР}|txt_{iTXT}$, компетенций переводчика относительно языковой пары $\mathbf{К}_{\mathbf{пер}i\mathbf{ПЕР}}$ и специализации переводчика относительно домена приложения текста $\mathbf{С}_{\mathbf{пер}i\mathbf{ПЕР},\ \mathbf{дп}i\mathbf{ДП}}$ формируется значение значимости $w_{\text{оц}k}$, множество нормированных значений $w_{\text{оц}k}$ значимости формирует матрицу значимости оценок сложности $W_{\text{оц}}$ 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 $C_{\text{оц}i}$ размерностью $1 \times k$, где k – общее число оценок.

Шаг 4. На основании оценок $C_{\text{оц}i}$ и значимости $W_{\text{оц}}$ формируется уравнение поиска теоретического результирующего фактора, т.е. качества перевода $\widehat{КП}$:

$$\widehat{\text{КП}}_i = w_0 + w_{\text{оц}1}\text{С}_{\text{оц}i1} + w_{\text{оц}2}\text{С}_{\text{оц}i2} + \cdots + w_{\text{оц}k}\text{С}_{\text{оц}ik}. \quad (4)$$

Для системы МП веса значимости оценок рассчитываются на основании тренировочных данных с использованием численного метода наименьших квадратов [12], при котором минимизируется сумма квадратов отклонений эмпирических значений результативного признака от теоретических, полученных по уравнению (4):

$$S(w) = \sum_{i=1}^{R}(\text{ОК}_i - \widehat{\text{ОК}}_i(\text{С}_{\text{оц}i}, w))^2,$$

$$S(w) = \sum_{i=1}^{R}(\text{ОК}_i - w_0 + w_{\text{оц}1}\text{С}_{\text{оц}i1} + w_{\text{оц}2}\text{С}_{\text{оц}i2} + \cdots + w_{\text{оц}k}\text{С}_{\text{оц}ik})^2 \to min, \quad (5)$$

где R – объем тренировочной выборки.

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

$$\sum_{i=1}^{R}(\text{ОК}_i - \widehat{\text{ОК}}_i(\text{С}_{\text{оц}i}, w))\frac{\partial \widehat{\text{ОК}}_r(\text{С}_{\text{оц}i}, w)}{\partial w} = 0. \quad (6)$$

Получаем систему нормальных уравнений с k неизвестными:

$$\begin{cases} \sum \text{ОК} = Rw_0 + w_{\text{оц}1}\sum \text{С}_{\text{оц}1} + w_{\text{оц}2}\sum \text{С}_{\text{оц}2} + \cdots + w_{\text{оц}k}\sum \text{С}_{\text{оц}k} \\ \sum \text{ОК}\cdot\text{С}_{\text{оц}1} = w_0\sum \text{С}_{\text{оц}1} + w_{\text{оц}1}\sum \text{С}_{\text{оц}1}{}^2 + w_{\text{оц}2}\sum \text{С}_{\text{оц}2}\text{С}_{\text{оц}1} + \cdots + w_{\text{оц}k}\sum \text{С}_{\text{оц}k}\text{С}_{\text{оц}1}. \\ \cdots \\ \sum \text{ОК}\cdot\text{С}_{\text{оц}k} = w_0\sum \text{С}_{\text{оц}k} + w_{\text{оц}1}\sum \text{С}_{\text{оц}1}\text{С}_{\text{оц}k} + w_{\text{оц}2}\sum \text{С}_{\text{оц}2}\text{С}_{\text{оц}k} + \cdots + w_{\text{оц}k}\sum \text{С}_{\text{оц}k}{}^2 \end{cases}$$

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

$$W_{\text{оц}} = (\text{С}_{\text{оц}}{}^T\cdot\text{С}_{\text{оц}})^{-1}\cdot\text{С}_{\text{оц}}{}^T\cdot\text{ОК} = (\frac{1}{R}\text{С}_{\text{оц}}{}^T\cdot\text{С}_{\text{оц}})^{-1}\frac{1}{R}\text{С}_{\text{оц}}{}^T\cdot\text{ОК}. \quad (7)$$

Шаг 5. Для каждого *i-го* фрагмента текста рассчитывается вероятность получения переведенного текста на таком уровне качества, который определяется требованиями $\text{ТР}|txt_{iTXT}$, применив к уравнению (4) логит-преобразование:

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

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

$$\text{СлЗП}_i = \frac{1}{p_i}. \quad (9)$$

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

$$\text{СлЗП}_{txti\text{ТХТ}} = \max \text{СлЗП}_i \quad (10)$$

В зависимости от значения $\text{СлЗП}_{txti\text{ТХТ}}$ определяется стратегия дальнейшей обработки текста, в том числе необходимость применять оптимизационное предредактирование.

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

Опишем условия задачи, пусть:

1) $txt'_{i\text{ТХТ}}$ – текст на языке $\text{яз}_{\text{вх}}$, созданный системой автоматического оптимизационного предредактирования, такой, при котором $\text{СМ}'_i \to \text{СМ}_i$ и $txt'_{i\text{ТХТ}} \neq txt_{i\text{ТХТ}}$, где СМ – смысл или упорядоченный набор семантических единиц, описываемый текстом: $\text{СМ} = \{(ce_0, ce_1, \dots, ce_{n\text{см}}): ce_{i\text{см}} \in \text{СЕ}\}$, $\text{СМ}'_i$ и СМ_i – смыслы $txt'_{i\text{ТХТ}}$ и $txt_{i\text{ТХТ}}$, соответственно;
2) $TXT_{i\text{src}}$ – множество всех возможных вариантов предредактированного текста, т.е. выражения смысла СМ_i текста $txt_{i\text{ТХТ}}$ на языке $\text{яз}_{\text{вх}}$*:*
3) $TXT_{i\text{src}} = \{txt_0, txt_1, \dots, txt_k\}$,
4) где k – обще число вариантов предредактированного текста, причем $txt_{i\text{ТХТ}}$, $txt'_{i\text{ТХТ}} \in TXT_{i\text{src}}$*;*
5) $\text{мСлЗП}_{i\text{src}}$ – множество оценок сложности задачи перевода вариантов предредактирования текста $txt_{i\text{ТХТ}}$ в соответствии с компетентностью $\mathbf{К_{перлПЕР}}$ и специализацией $\mathbf{С_{перлПЕР,\ дплДП}}$ системы МП $\text{пер}_{i\text{ПЕР}}$ для всех возможных вариантов предредактированного текста $TXT_{i\text{src}}$*:*
6) $\text{мСлЗП}_{i\text{src}} = \{\text{СлЗП}_0, \text{СлЗП}_1, \dots, \text{СлЗП}_k\}$;

7) Каждому варианту предредактированного текста соответствует одна оценка сложности задачи перевода для системы МП $\text{пер}_{i\text{ПЕР}}$, то есть множества $TXT_{i\text{src}}$ и $\text{мСлЗП}_{i\text{src}}$ биективны: $TXT_{i\text{src}} \leftrightarrow \text{мСлЗП}_{i\text{src}}$;
8) $\text{СлЗП}_k \in \text{мСлЗП}_{i\text{src}}$ – оценка сложности задачи перевода варианта предредактированного текста $txt'_{i\text{ТХТ}}$ для системы МП $\text{пер}_{i\text{ПЕР}}$;
9) $[min\text{СлЗП};\ max\text{СлЗП}]$ – диапазон значений оценок сложности задачи перевода $\text{мСлЗП}_{i\text{src}}$;
10) $\text{СлЗП}_{\text{доп}}$ – максимально допустимое значение критерия «Низкая сложность задачи перевода» при допущении, что чем ниже значение СлЗП_k, тем лучше;
11) нСлЗП_i – нечёткое подмножество множества $\text{мСлЗП}_{i\text{src}}$, определяющее принадлежность элементов множества $\text{мСлЗП}_{i\text{src}}$ и соответствующих элементов множества $TXT_{i\text{src}}$ классу «Низкая сложность задачи перевода»;
12) $\text{нСлЗП}_i = \{(\text{СлЗП},\ \mu_{\text{нСлЗП}i}(\text{СлЗП}))|\text{СлЗП} \in \text{мСлЗП}_{i\text{src}}\}$;
13) $\mu_{\text{нСлЗП}i}(\text{СлЗП})$ – функция принадлежности, указывающая в какой степени текст txt с оценкой СлЗП принадлежит нечеткому множеству нСлЗП_i;
14) $\mu_{\text{нСлЗП}i}(\text{СлЗП}) \in [0;\ 1]$ и имеет вид логистической кривой:

$$\mu_{\text{нСлЗП}i}\ (\text{СлЗП}) = \frac{1}{1+e^{\left(\frac{\text{СлЗП}-\text{СлЗП}_{\text{доп}}}{|min\text{СлЗП}-\text{СлЗП}_{\text{доп}}|}\right)2\pi}}. \quad (11)$$

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

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

$$F_{\text{АОПР}}(\Psi, \mu_{\text{нСлЗП}i}(\text{СлЗП})) = \prod_{l=1}^{L} \text{F}_{\text{АОПР}}(\mu_{\text{нСлЗП}i}(\text{СлЗП})_l) =$$
$$= P_\Psi(\mu_0 = \max \mu_{\text{нСлЗП}i}(\text{СлЗП})_0) \cdot \dots \cdot P_\Psi(\mu_l = \max \mu_{\text{нСлЗП}i}(\text{СлЗП})_l) = \quad (12)$$
$$= P_\Psi(\mu_0 = \max \mu_{\text{нСлЗП}i}(\text{СлЗП})_0), \dots, \mu_l = \max \mu_{\text{нСлЗП}i}(\text{СлЗП})_l),$$

где l – номер экземпляра в обучающей выборке объемом L.

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

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

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

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

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

Решение поставленной задачи оптимизации выполняется методом градиентного спуска (подъема) [13]. Для этого необходимо найти градиент логарифмической функции правдоподобия $L_{\text{АОПР}}(\Psi, \mu_{\text{нСлЗП}i}(\text{СлЗП}))$ – вектор, который показывает направление возрастания функции.

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

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

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

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

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

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 $\nabla L_{\text{АОПР}}(\Psi)$:

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

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:

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

Используя описанную математическую модель, создадим модель автоматического оптимизационного предредактирования.

4. Обучающие данные для модели оценки сложности задачи перевода. Для обучения модели оценки сложности задачи перевода заданного текста $txt_{i\text{ТХТ}}$ с языка $\text{яз}_{\text{вх}}$ на язык $\text{яз}_{\text{вых}}$ переводчиком $\text{пер}_{i\text{ПЕР}}$ в соответствии с формализованными требованиями к переводу $\text{ТР}|txt_{i\text{ТХТ}}$ требуются исходные данные в виде корпуса параллельных текстов следующей структуры:

$TranslatorExpCor$: $[src;\ trg;\ ref]$, где src – это оригинал, т.е. текст на языке $\text{яз}_{\text{вх}}$, trg – это перевод (текст на языке $\text{яз}_{\text{вых}}$), выполненный переводчиком или системой МП $\text{пер}_{i\text{ПЕР}}$; ref – это контрольный перевод (текст на языке $\text{яз}_{\text{вых}}$), т.е. проверенный эталон.

Путем обработки корпуса $TranslatorExpCor$ модулями оценки качества МП и препроцессинга текстовых данных для взвешенной оценки параметров русскоязычного текста формируется база данных структурного анализа предложений, содержащая расчет вещественных параметров по морфологическим, синтаксическим, лексическим и прочим признакам [14].

5. Обучающие данные для модели автоматического оптимизационного предредактирования. Для обучения модели, которая будет преобразовывать текст на языке $\text{яз}_{\text{вх}}$ в текст требуемой структуры для повышения качества перевода, необходимо создать корпус тренировочных текстов в паре $\text{яз}_{\text{вх}} - \text{яз}_{\text{вх}}$. 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 $\text{яз}_{\text{вх}}$, tgt_ref – это перевод (текст на языке $\text{яз}_{\text{вых}}$).

Методика сбора корпуса обучающих текстов для модели оптимизационного предредактирования включает следующие шаги:

1. Настраиваем системы МП MT: tgt – src, MT: src – tgt.

При помощи системы MT: tgt – src переводим текст tgt_ref на язык $\text{яз}_{\text{вх}}$, получим массив текстовых данных pre_src (массив условно предредактированных текстов).

При помощи системы MT: src – tgt переводим текст src_ref на язык $\text{яз}_{\text{вых}}$, получим массив текстовых данных tgt1 (src_ref → tgt1).

При помощи системы MT: src – tgt переводим текст pre_src на язык $\text{яз}_{\text{вых}}$, получим массив текстовых данных tgt2 (src_ref → tgt1).

Оцениваем качество выполненного перевода на язык $\text{яз}_{\text{вых}}$ tgt1 и tgt2 относительно эталона tgt_ref, получаем массивы оценок QC_score(tgt1) и QC_score(tgt2).

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 $\Delta\text{QC\_score}_i$ является условно значимой $d_{\max}$ для выбранного типа оценки:

$$\text{TrainCor} = \{(\text{src}_{\text{ref}_i};\ \text{pre}_{\text{src}_i}) : \exists\ (\text{tgt1}_i, \text{tgt2}_i) | \text{QC\_score}(\text{tgt2}_i) > \\ > \text{QC\_score}(\text{tgt1}_i)\ \&\ \Delta\text{QC\_score}_i \geq d_{\max}\} \quad (19)$$

Полученный корпус обучающих текстов TrainCor объемом выборки L будем использовать для обучения языковой модели LM: $\text{src} - \text{pre\_src}$ для решения задачи автоматического оптимизационного редактирования текстов на языке $\text{яз}_{\text{вх}}$.

6. Оценка качества выполненного перевода. Для реализации программного комплекса определим критерий качества МП QC_score. Традиционно, для оценки качества перевода используются алгоритмы, которые сравнивают выполненный перевод с одним или несколькими эталонными переводами при помощи числовой метрики. В рамках исследования была выбрана метрика hLEPOR [15], которая имеет наивысший балл корреляции Пирсона с человеческими суждениями по языковой паре английский-русский.

$hLEPOR$ – гармоническое среднее между штрафом за длину перевода в сравнении с эталоном, штрафом за различие в позициях и степенью перекрытия n-грамм; рассчитывается в диапазоне от 0 до 1, где 0 – полое несовпадение; 1 – полное совпадение:

$$hLEPOR = \frac{(LP \cdot PP \cdot OR)^{\frac{1}{3}}}{(\frac{1}{LP} + \frac{1}{PP} + \frac{1}{OR})^{\frac{1}{3}}} \quad (20)$$

где LP — штраф за длину, который учитывает различие в длине между переводом и эталонным переводом; PP — штраф за различие в позициях, который учитывает различие в расположении слов в переводе и эталоне; OR — степень перекрытия, которая учитывает совпадение n-грамм в переводе и эталоне.

Для расчета каждого из компонентов используются следующие формулы:

$$LP = f(x) = \begin{cases} e^{1-\frac{ref}{cand}}, & cand > ref \\ 1, & cand \leq ref \end{cases}, \quad (21)$$

где $cand$ — длина переведенного текста, а ref — длина эталонного перевода;

$$PP = \frac{1}{1+p_{diff}}, \quad (22)$$

где p_{diff} — среднее арифметическое различий позиций одинаковых n-грамм в переводе и эталоне;

$$OR = BP \cdot ng_{prec} \cdot (1 - pen), \quad (23)$$

где BP — штраф за слишком короткий перевод, pen — дополнительный штраф за ошибки, ng_{prec} — точность n-грамм, которая рассчитывается как отношение количества совпадающих n-грамм в переводе и эталоне к общему количеству n-грамм в переводе.

6. Программный комплекс оптимизационного предредактирования русскоязычных текстов. Опираясь на описанные математические модели и методики, была разработана архитектура программного комплекса, который состоит из трех основных подсистем: подсистемы тренировки языковой модели, подсистемы оценки сложности задачи перевода, подсистемы оптимизационного предредактирования русскоязычного текста и генерации МП на английский язык. При реализации приняты 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.
   Схематично архитектура программного комплекса представлена на рисунке 2.

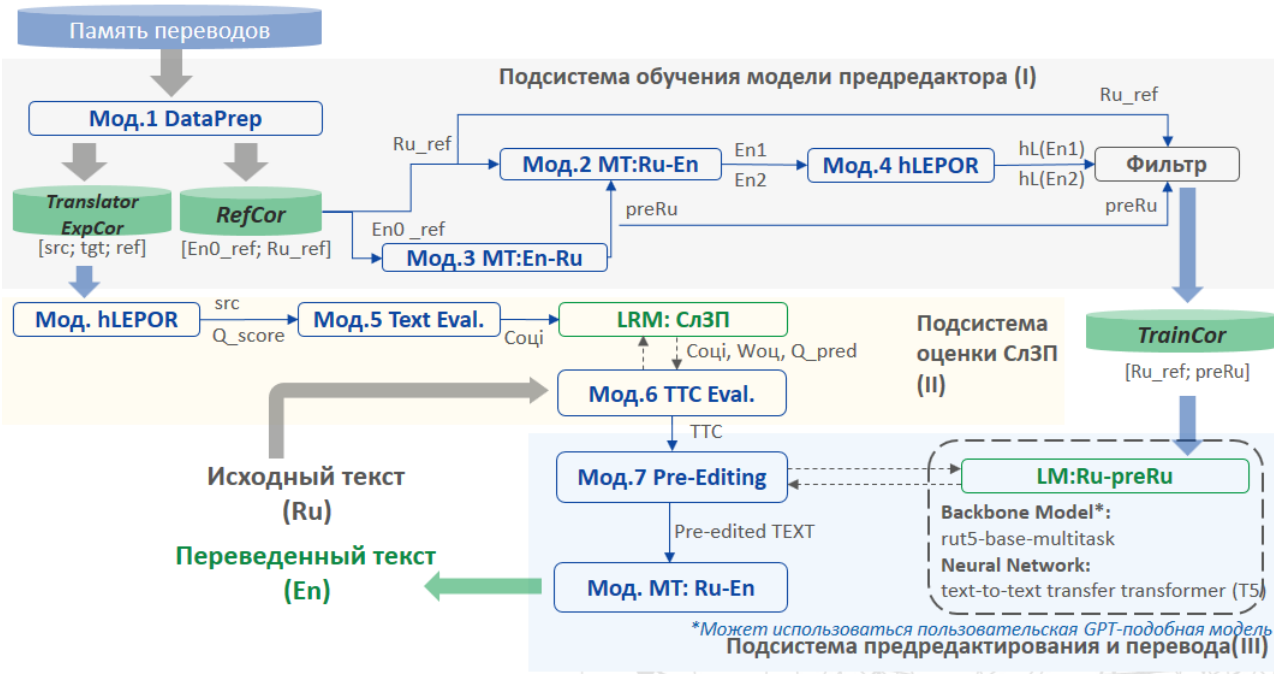

Рис. 2. Архитектура программного комплекса оптимизационного предредактирования русскоязычных текстов

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), представленной командой Google в 2020 г., дообученной на корпусе параллельных русскоязычных текстов подзадачу перефразирования [16]; модуля генерации МП с русского языка на английский язык на основе модели *Helsinki-NLP/opus-mt-ru-en*.

7. Тестирование программного комплекса. Программный комплекс [17] 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», созданного и функционирующего на базе ВЦ ДВО РАН – обособленном подразделении ХФИЦ ДВО РАН. Тип используемой ЭВМ: компьютер с архитектурой x86, x86_64.

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 ~140 000 экземпляров и TranslatorExpCor объемом ~90 000 экземпляров. После обработки корпуса RefCor в подсистеме тренировки языковой модели оптимизационного предредактирования русскоязычных текстов в корпус TrainCor вошло ~84 000 экземпляров. Тестовая выборка для оценки работы системы TestCor составила ~17 000 экземпляров.

Тестирование программного комплекса включало следующие этапы:

1. Машинный перевод тестовой выборки и оценка его качества алгоритмом hLEPOR → получение оценки $hLEPOR(En1)$.
2. Оценка сложности задачи перевода тестовой выборки $\text{СлЗП}(Ru)$ → применение оптимизационного предредактирования только к тем семплам, для которых $\text{СлЗП}(Ru) > \text{СлЗПдоп}$ → оценка СлЗП после оптимизационного предредактирования $\text{СлЗП}(preRu)$ → машинный перевод → получение оценки $hLEPOR\ (TTC/PE)$.
3. Сравнение и анализ полученных результатов.

В рамках тестирования принято, что минимально допустимая сложность задачи перевода $\text{СлЗП}_{\text{доп}} = 1{,}43$. Так как сложность задачи перевода обратно пропорциональна вероятности получения перевода требуемого качества, при $\text{СлЗП}_{\text{доп}} = 1{,}43$ данная вероятность составляет 0,7.

С использованием оценки сложности задачи перевода из тестовой выборки было отобрано 5440 экземпляров для оптимизационного предредактирования (32,58% тестовой выборки).

Примеры оптимизационного предредактирования и его влияния на сложность задачи перевода для системы МП представлены в таблице 1. Результаты перевода текстов после оптимизационного предредактирования на английский язык представлены в таблице 2.

Таблица 1. Оптимизационное предредактирование русскоязычных текстов

Экз.	**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 если такое оборудование отсутствует, в центральном шкафу предусматривают установку контроллера автоматического ввода резерва.	Оборудование должно быть способно управлять двумя фидерами, в случае отсутствия такого оборудования в центральном шкафу должен быть установлен переключатель ввода резерва.	-0,444

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) об отклонениях от плана работы.	-0,271

Таблица 2. Результаты машинного перевода на английский язык

Экз.	**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 3.

Таблица 3. 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

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.

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

Использование полученных результатов исследования и внедрение программного комплекса в работу агентства переводов «ФИАС-Амур» позволило повысить эффективность использования систем МП и производительность труда редакторов переводов, а также оптимизировать затраты на оказание услуг перевода узкоспециальной технической документации.

В ходе внедрения 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 увеличилась с 3,8 до 4,3 стандартных страницы в час. Таким образом, внедрение программного комплекса позволило увеличить производительность редакторов переводов на 13,16%, что является значимым показателем при переводе большого объема документации в условиях дефицита квалифицированных кадров.

Описанные модели и методы 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 инструмент объективной оценки исполнителей в рамках поставленной задачи на перевод.

Программный комплекс может быть внедрен в компаниях, генерирующих от 1000 страниц перевода в месяц, предоставив инструментарий повышения качества перевода, в том числе для редакторов без знания языка перевода.

Заключение. В ходе работы 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; 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.

Программный комплекс представляет собой ансамбль алгоритмов и моделей, включая модель классификации и генеративную модель русского языка, каждая из которых имеет потенциал к доработке с целью повышения точности, что позволит улучшить эффективность программного комплекса в целом.

Результаты исследования подтверждают эффективность применения оптимизационного предредактирования русскоязычных узкоспециальных текстов с целью повышения качества МП на английский язык. Разработанный авторами программный комплекс оптимизационного предредактирования имеет потенциал к доработке и повышении точности. Интеграция программного комплекса в контур автоматизации процессов перевода технической документации позволяет снизить затраты на постредактирование МП и организацию переводческих процессов.

Благодарности. Работа выполнена в рамках соглашения о стратегическом научно-технологическом сотрудничестве между ФГБОУ ВО «КнАГУ» (г. Комсомольск-на-Амуре), Центром обработки и хранения научных данных ДВО РАН на базе ВЦ ДВО РАН – обособленного подразделения ХФИЦ ДВО РАН (г. Хабаровск) и ООО «Агентство переводов «ФИАС-Амур» (г. Комсомольск-на-Амуре).

Список источников

1. Quinci C., Pontrandolfo G. Testing neural machine translation against different levels of specialization. Transkom, 2023, vol.1, pp.174-209.
2. Canfora C., Ottmann A. Risks in neural machine translation. Translation spaces, 2020, vol. 9(1), pp. 58–77.
3. Kumar V., Kulkarni A., Singh P., Ramakrishnan G. A machine assisted human translation system for technical documents. Miami MT Summit XV, 2015, vol.2, pp. 259-272.
4. 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, pp. 64-74.
5. 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.
6. Taufik A. Pre-editing of Google neural machine translation. Journal of English language and culture, 2020, vol. 10, no. 2, pp. 64-74.

7. O'Brien, S. Controlling controlled English: an analytical of several controlled language rule sets. Proceedings of EAMT-CLAW, Dublin, Ireland, 2003, pp. 105-114.
8. Оборнева И. В. Автоматизация оценки качества восприятия текста / И. В. Оборнева // ВЕСТНИК Московского городского педагогического университета, 2015. – №2(5). – С. 221–233.
9. Дмитриева А.Д. Квантитативное исследование стратегий упрощения на материале адаптированных текстов для изучающих РКИ / А.Д. Дмитриева, А.Н. Лапошина, М.Ю. Лебедева // Компьютерная лингвистика и интеллектуальные технологии: по материалам международной конференции «Диалог», 2021. – С. 191-204.
10. 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, pp. 1-15.
11. Barber D. Bayesian Reasoning and machine learning. Cambridge: Cambridge university press, 2012, DOI:10.1017/CBO9780511804779.
12. MacKay D.J.C. Information theory, inference, and learning algorithms. Cambridge: Cambridge University Press, 2003, DOI: 10.2277/0521642981
13. Bishop, C.M. Pattern Recognition and Machine Learning. Springer, Berlin, 2006.
14. Животова А.А. 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.
15. Li-Feng Han A., Wong D.F., Chao L.S. et al. Language-independent model for machine translation evaluation with reinforced factors. Proceedings of machine translation Summit XIV: Posters, Nice, France, 2013, pp. 215-222.
16. Дале Д. Многозадачная модель T5 для русского языка. – URL: https://habr.com/ru/articles/581932/ (дата обращения: 20.08.2023).
17. Свидетельство о государственной регистрации программы для ЭВМ № 2023682260 Российская Федерация. Программный комплекс для предредактирования и машинного перевода узкоспециальных русскоязычных текстов на английский язык: № 2023680875: заявл. 09.10.2023: опубл. 24.10.2023 / А. А. Животова, В. Д. Бердоносов; заявитель Животова А.А. – 1 с.

Животова Алена Анатольевна. *Аспирант, ассистент кафедры Прикладная математика факультета компьютерных технологий Комсомольского-на-Амуре государственного университета, ORCID: 0000-0003-1037-5503, SPIN: 3142-9290, РИНЦ AuthorID: 771774, zhivotova.aa@gmail.com, г. Комсомольск-на-Амуре, пр. Ленина, д. 27.*

Бердоносов Виктор Дмитриевич. *К.т.н., доцент кафедры Прикладная математика факультета компьютерных технологий Комсомольского-на-Амуре государственного университета, ORCID: 0000-0003-4093-779X, SPIN: 8730-2226, РИНЦ AuthorID: 644399, berd1946@gmail.com, г. Комсомольск-на-Амуре, пр. Ленина, д. 27.*

UDC 004.9:519.768
DOI:10.25729/ESI.2024.34.2.016

Optimizational pre-editing of highly specialized Russian-language texts for its machine translation into English

Alena A. Zhivotova, Victor D. Berdonosov

Komsomolsk-na-Amure State University,
Russia, Komsomolsk-na-Amure, *zhivotova.aa@gmail.com*

Abstract. The authors study the process of text translation, particularly the method of optimizing pre-editing as a way to improve the quality of machine translation into English for Russian-language highly specialized texts. The paper considers the mathematical model of translation process and machine translation task formulation, proposes a new theory for probabilistic estimation of translation task complexity, provides the formulation and solution of optimizing pre-editing task, describes data preparation methodology for training automatic optimizing pre-editing model. As the research result the software package for optimizing pre-editing of Russian-language texts is developed. The software package has been developed using resources of the Center for Scientific Equipment Collective Use "Center for Processing and Storage of Scientific Data of the Far Eastern Branch of the Russian Academy of Sciences". Data for models training and validation are provided by Translation Agency FIAS-Amur Co., Ltd. Software package testing has proved the effectiveness of the proposed methods for improving the quality of machine translation of highly specialized Russian-language texts into English.

Keywords: machine translation, optimizational pre-editing, translation task complexity, translation quality

Acknowledgements: the research was conducted within the agreement on strategic scientific and technological cooperation between Komsomolsk-na-Amure State University (Komsomolsk-na-Amure), Center for Processing and Storage of Scientific Data of the Far Eastern Branch of the Russian Academy of Sciences (Khabarovsk), and Translation Agency FIAS-Amur Co., Ltd. (Komsomolsk-na-Amure).

References

1. Quinci C., Pontrandolfo G. Testing neural machine translation against different levels of specialization. Transkom, 2023, vol.1, pp.174-209.
2. Canfora C., Ottmann A. Risks in neural machine translation. Translation spaces, 2020, vol. 9(1), pp. 58–77.
3. Kumar V., Kulkarni A., Singh P., Ramakrishnan G. A machine assisted human translation system for technical documents. Miami MT Summit XV, 2015, vol.2, pp. 259-272.
4. 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, pp. 64-74.
5. 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.
6. Taufik A. Pre-editing of Google neural machine translation. Journal of English language and culture, 2020, vol. 10., no. 2, pp. 64-74.
7. O'Brien, S. Controlling controlled English: an analytical of several controlled language rule sets. Proceedings of EAMT-CLAW, Dublin, Ireland, 2003, pp. 105-114.
8. Oborneva, I. V. Avtomatizaciya ocenki kachestva vospriyatiya teksta [Automation of text perception quality assessment]. VESTNIK Moskovskogo gorodskogo pedagogicheskogo universiteta [Herald of Moscow City Pedagogical University], 2015, no.2(5), pp. 221-233.
9. Dmitrieva A.D., Laposhina A., Lebedeva M. Kvantitativnoe issledovanie strategij uproshheniya na materiale adaptirovanny`x tekstov dlya izuchayushhix RKI [A Quantitative Study of Simplification Strategies in Adapted Texts for L2 Learners of Russian]. Komp`yuternaya lingvistika i intellektual`ny`e texnologii: po materialam mezhdunarodnoj konferencii "Dialog" [Computer Linguistics and Intelligent Technologies: Proceedings of the International Conference "Dialog"], 2021, pp. 191-204.
10. 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, pp. 1-15.
11. Barber D. Bayesian Reasoning and machine learning. Cambridge: Cambridge university press, 2012, DOI:10.1017/CBO9780511804779.
12. MacKay D.J.C. Information theory, inference, and learning algorithms. Cambridge: Cambridge University Press, 2003, DOI: 10.2277/0521642981
13. Bishop, C.M. Pattern Recognition and Machine Learning. Springer, Berlin, 2006.
14. Zhivotova A.A., Berdonosov V.D. Regressionnyj analiz korrelyacii kachestva mashinnogo perevoda i parametrov iskhodnogo teksta [Regression analysis of the correlation between machine translation quality and source text features]. Informatika i sistemy upravleniya [Information science and control systems], 2023, no.2(76), pp.121-133.
15. Li-Feng Han A., Wong D. F., Chao L.S. et al. Language-independent model for machine translation evaluation with reinforced factors. Proceedings of machine translation Summit XIV: Posters, Nice, France, 2013, pp. 215-222.
16. Dale, D. Mnogozadachnaya model` T5 dlya russkogo yazy`ka [T5 multitasking model for Russian language]. Available at: https://habr.com/ru/articles/581932/ (accessed: 08/20/2023).
17. Zhivotova A.A., Berdonosov V.D. Programmny`j kompleks dlya predredaktirovaniya i mashinnogo perevoda uzkospecial`ny`x russkoyazy`ch-ny`x tekstov na anglijskij yazy`k [Program complex for pre-editing and machine translation of highly specialized Russian-language texts into English]. Software Registration Certificate of the Russian Federation no. 2023680875 (2023).

Alena Anatolievna Zhivotova. *Postgraduate student, assistant of Applied Mathematics department, faculty of computer technologies, Komsomolsk-na-Amure state university, ORCID: 0000-0003-1037-5503, SPIN: 3142-9290, AuthorID: 771774, zhivotova.aa@gmail.com, Russia, Komsomolsk-na-Amure, 27 Lenin Ave.*

Victor Dmitrievich Berdonosov. *PhD in technical sciences, associate professor of Applied Mathematics department, faculty of computer technologies, Komsomolsk-na-Amure state university, ORCID: 0000-0003-4093-779X, SPIN: 8730-2226, AuthorID: 644399, berd1946@gmail.com, Russia, Komsomolsk-na-Amure, 27 Lenin Ave.*

Статья поступила в редакцию 08.12.2023; одобрена после рецензирования 29.03.2024; принята к публикации 06.06.2024.

The article was submitted 12/08/2023; approved after reviewing 03/29/2024; accepted for publication 06/06/2024.