УДК 81`322.4

Епимахова Александра Сергеевна
кандидат филологических наук,
доцент базовой кафедры технологий
и автоматизации перевода в бюро
переводов «АКМ-Вест»
Северный (Арктический)
федеральный университет имени
М. В. Ломоносова
г. Архангельск, Россия

Aleksandra Epimakhova
PhD in Philology,
Associate Professor of the Department
of Translation Technology and Practice
at AKM-West
Northern (Arctic) Federal University
Arkhangelsk, Russia
a.epimahova@narfu.ru

Коканова Елена Сергеевна
кандидат филологических наук,
заведующий базовой кафедрой
технологий и автоматизации перевода
в бюро переводов «АКМ-Вест»
Северный (Арктический)
федеральный университет
имени М.В. Ломоносова
г. Архангельск, Россия

Elena Kokanova
PhD in Philology,
Head of the Department of Translation
Technology and Practice at AKM-West
Northern (Arctic) Federal University
Arkhangelsk, Russia
e.s.kokanova @narfu.ru

АВТОМАТИЧЕСКАЯ ОЦЕНКА КАЧЕСТВА МАШИННОГО ПЕРЕВОДА В ЯЗЫКОВОЙ ПАРЕ С МАЛОРЕСУРСНЫМ ЯЗЫКОМ

Использование метрик автоматической оценки качества стало неотъемлемой частью разработки систем машинного перевода. Выбор метрик определяется такими факторами, как языковая пара и направление перевода, возможность интуитивной интерпретации полученных результатов, популярность метрики среди разработчиков и исследователей. В статье проводится экспериментальное исследование по разработке модуля оценки качества машинного перевода для языковой пары ненецкий-русский. В результате исследования для автоматической оценки качества выбраны классические метрики BLEU, sacreBLEU, TER, CharacTER, RIBES, chrF++ для обоих направлений перевода и нейросетевая метрика BERTScore только для направления ненецкий-русский. Ограничения связаны с тем, что необходимо использовать метрики, либо рассчитываемые независимо от конкретного языка, либо поддерживающие данный язык. Число метрик, поддерживающих русский язык, ограничено. Ненецкий язык относится к малоресурсным и не поддерживается метриками. Использование нескольких метрик позволяет получить более комплексную оценку.

Ключевые слова: машинный перевод; автоматическая оценка качества машинного перевода; метрики автоматической оценки качества машинного перевода; малоресурсный язык; ненецкий язык; языковая пара ненецкий-русский.

AUTOMATIC MACHINE TRANSLATION EVALUATION FOR A LOW-RESOURCE LANGUAGE

Automatic evaluation metrics are an integral part of machine translation systems development. The choice of metrics is determined by the source and target languages, possibility of intuitive interpretation of the scores, and the popularity of the metric among developers and

researchers. The paper presents an experimental study on the development of a machine translation evaluation module for the Nenets-Russian language pair. As a result of the study, the traditional metrics BLEU, sacreBLEU, TER, CharacTER, RIBES, chrF++ were selected for automatic evaluation in both translation directions. BERTScore neural network metric was selected only for the Nenets-Russian direction. The chosen metrics should either be calculated independently of a specific language, or support the given languages. However, the number of metrics supporting Russian is limited. As for the Nenets language, it is low-resourced and therefore not supported by any metric. Using multiple metrics permits to get a more comprehensive assessment.

Key words: machine translation; automatic machine translation evaluation; automatic machine translation evaluation metrics; low-resource language; Nenets language; Nenets-Russian language pair.

В настоящее время автоматическая оценка качества машинного перевода (МП) с использованием метрик стала не просто альтернативой экспертной оценке, призванной снизить временные и финансовые затраты. Она включена в цикл разработки систем МП и позволяет в числовой форме выразить оценку качества МП (предполагается, что эта оценка соотносится с экспертной оценкой) [1]. Кроме того, автоматическая оценка позволяет сопоставлять системы МП между собой, определять, какая из них покажет лучший результат для определенного типа текста [2, p. 7297]. Важными требованиями к метрикам автоматической оценки качества МП, помимо возможности различать по качеству разные системы МП, а также корреляции с результатами экспертной оценки МП, являются возможность интуитивной интерпретации полученных результатов, ясность и применимость к разным языкам перевода [3]. Подключение экспертной оценки позволяет подкрепить автоматическую оценку [4, с. 108].

Выбор метрик важен при разработке и использовании систем МП, а также публикации полученных результатов [5], так как разные метрики позволяют не только учесть особенности языковой пары, но и сделать результаты более доступными благодаря использованию распространенных метрик. При выборе метрик автоматической оценки МП учитываются разные факторы.

Факторы, определяющие выбор метрик автоматической оценки МП

При выборе метрики важным фактором является возможность ее расчета с использованием имеющихся в открытом доступе библиотек, таких как NLTK, sacrebleu, evaluate. Из множества разработанных метрик лишь самые популярные реализованы в библиотеках. Они позволяют с использованием языка программирования Python загрузить на вход текст или предложение, передать параметры оценки и получить результат.

Результаты оценки по одной и той же метрике могут иметь разное числовое выражение. Например, результаты метрики BLEU (BiLingual Evaluation Understudy [6]), рассчитываемой на основе пересечения n-грамм, зависят от предварительной подготовки текста (она может включать удаление знаков препинания, приведение к нижнему регистру и т. д.), используемого механизма токенизации, количества предложений-эталонов и длины

n-грамм. Поскольку эти параметры не всегда представлены в публикациях, следует с осторожностью сравнивать оценки BLEU из разных статей [7]. При этом в 2021 году BLEU использовалась в 98,8 % публикаций, причем в 74,3 % случаев без привлечения других метрик [2, p. 7299]. Для стандартизации BLEU была создана SacreBLEU, имеющая встроенный токенизатор и использующая установленные параметры оценки (для поддерживаемых языков возможно подключение данных из набора экспертной оценки WMT, Workshop on Machine Translation) [8].

Метрики, для которых есть возможность расчета с использованием библиотек, могут поддерживать не все языки. Например, на основе BLEU построена метрика NIST [9], особенность которой состоит в том, что каждой n-грамме в МП и в эталонном переводе присваивается определенный вес, зависящий от частности элемента в языке перевода, и этот вес учитывается при расчете метрики. Метрика NIST, рассчитываемая с использованием библиотеки NLTK, реализована для английского языка. Метрика METEOR (Metric for Evaluation of Translation with Explicit ORdering) [10] также имеет в основе пересечение n-грамм. Ее ключевое отличие от BLEU состоит в том, что учитываются не только совпадения полных слов, но также наличие однокоренных слов и синонимов. Поэтому метрика METEOR реализована в библиотеке NLTK с подключением по умолчанию тезауруса WordNet [11], т. е. имеет смысл использовать ее только если языком перевода является английский. Следовательно, преимущество BLEU состоит в ее универсальности и удобстве использования для разных языков.

Среди классических метрик, которые не требуют поддержки конкретного языка, следует назвать RIBES, ROUGE, TER, CharacTER, chrF++.

RIBES (Rank-based Intuitive Bilingual Evaluation Score) [12] позволяет учесть позиции слов в предложении, что важно для оценки МП, когда порядок слов в языках сильно различается.

ROUGE (Recall-Oriented Understudy for Gisting Evaluation) [13] изначально разработана для оценивания результата суммаризации текста (сопоставление аннотации с эталонной аннотацией) и позволяет оценить количество элементов эталонного текста, представленных в оцениваемом.

Метрика TER (Translation Edit Rate) [14] основана на подсчете минимального количества правок, которое необходимо для того, чтобы получить перевод, наиболее близкий к эталонному (итоговый показатель – отношение числа требуемых правок к длине предложения, усредненной по всем эталонам). CharacTER (Translation Edit Rate on Character Level) [15] оценивает необходимое количество правки, но рассчитывается его на основе символов, а не слов, как TER (что актуально для языков с богатой морфологией, так как иногда требуется исправление не целого слова, а только его части).

В основе метрики chrf++ (correlations with human rankings for 6-gram F1-score) [16, p. 612] лежит подсчет совпадений последовательно идущих символов, а также совпадения униграмм и биграмм.

Исследования показывают, что классические метрики уступают в корреляции с экспертной оценкой нейросетевым [17], хотя результаты оценки последних труднее поддаются интерпретации, так как логика ее получения остается неизвестной [18]. Однако нейросетевые метрики имеют в основе модели, обученные на данных для конкретных языков, поэтому они реализованы не для всех языков (прежде всего для английского).

Русский язык поддерживает метрика BERTScore [19]. BERTScore вычисляет близость двух предложений как сумму косинусных сходств между эмбеддингами слов в предложениях. BERTScore выдает комплекс из трех оценок: Precision (точность), Recall (полнота) и F1 (F-мера). Precision показывает, насколько для каждого токена МП есть близкий ему токен эталонного предложения; Recall показывает, насколько для каждого токена эталонного предложения есть близкий ему токен МП. F1 учитывает оба эти параметра и рассчитывается по формуле: F1 = 2 * (Precision * Recall) / (Precision + Recall). BERTScore использует 12 предобученных моделей эмбеддингов для разных языков и обучена на корпусе WMT18 metric evaluation dataset, который содержит МП 149 разных систем для 14 языковых пар, эталонные варианты перевода, а также результаты экспертной оценки WMT [19, p. 5].

Также русский язык поддерживает COMET (Crosslingual Optimized Metric for Evaluation of Translation) [20]. COMET – это нейросетевая метрика, построенная на PyTorch. При обучении модели использовались не только языковые данных, но и экспертные оценки переводов WMT. Особенность COMET состоит в том, что она получает на вход МП, эталонный перевод и текст оригинала. Следовательно, возможность применения метрики ограничена поддержкой обоих языков языковой пары.

Несмотря на более высокую точность нейросетевых метрик, преимущество традиционных метрик состоит в том, что многие из них не требуют привязки к конкретному языку. Это тем более важно, если в языковую пару входит малоресурсный язык, т. е. язык «с малым объемом электронных ресурсов, доступных для обработки» [21, с. 679]). Малоресурсные языки зачастую не имеют корпусов, словарей, электронных библиотек для обработки естественного языка.

Экспериментальное исследование

В рамках эксперимента по разработке системы МП для языковой пары «ненецкий-русский» были выбраны метрики, которые могут быть использованы для оценки качества МП в обоих направлениях.

При выборе учитываются следующие критерии: (1) расчет с использованием библиотек; (2) применимость к любому языку или поддержка языка перевода; (3) использование только эталонного перевода, без текста оригинала (в направлении ненецкий-русский); (4) оригинальность лежащих в основе критериев; (5) использование в публикациях. В результате собран набор метрик, учитывающих разные аспекты (совпадение n-грамм, объем правки, порядок слов и др.).

В результате из представленных выше метрик для направления перевода «ненецкий-русский» выбраны классические (BLEU, sacreBLEU, TER, CharacTER, RIBES, chrF++) и нейросетевая BERTScore. Они реализованы в библиотеках NLTK, evaluate и sacrebleu. COMET не может быть использована, так как она поддерживает русский язык, но не поддерживает ненецкий, требуя загрузки текста оригинала. Было принято решение исключить из набора метрик ROUGE, так как она изначально не предназначена для оценки МП; редко используется в современных публикациях, посвященных МП; выдает громоздкую цепочку оценок; при этом схожие параметры оцениваются chrF++. BLEU и sacreBLEU, TER и CharacTER по сути являются вариантами одной метрики, однако они часто используются в публикациях, поэтому целесообразно иметь оба варианта оценок. RIBES отличается от прочих метрик таким уникальным параметром как учет порядка слов, что важно для ненецкого и русского языков, имеющих разный синтаксис.

Для направления перевода «русский-ненецкий» могут быть использованы только классические метрики (выбраны BLEU, sacreBLEU, TER, CharacTER, RIBES, chrF++).

Для разработки алгоритма автоматической оценки (на уровне предложения) использован набор, содержащий 10 предложений на русском языке, которые имеют одно предложение-эталон. Оцениваемые предложения моделируют МП и представляют собой варианты перевода исходного предложения, в том числе содержащие ошибки.

Исходное предложение: *Хувы ялумдаңкаңна нисяв си''ми сиде.*

Эталонный перевод: *Отец разбудил меня на утренней заре.*

Оцениваемые предложения:

1. *Отец разбудил меня на утренней заре.*
2. *На утренней заре меня разбудил отец.*
3. *Папа разбудил меня на утренней заре.*
4. *Мой отец разбудил меня на утренней заре.*
5. *Отец разбудил меня утром на заре.*
6. *Отец разбудил меня на заре.*
7. *Отец не разбудил меня на утренней заре.*
8. *Мать разбудила меня на утренней заре.*
9. *У меня есть летний чум.*
10. *У нас есть летний чум.*

Предложения, выступающие в роли МП, имеют характеристики, которые позволяют увидеть особенности разных метрик: полное совпадение с эталоном (1), изменение порядка слов (2), замена знаменательного слова на синоним (3), добавление служебного слова, которое не меняет смысл предложения (4), использование однокоренного слова (5), удаление слова без изменения смысла предложения (6), добавление отрицания (7), замена слова на другое, используемое в схожих контекстах, с изменением смысла (8), совпадение только одного слова (9), полное несовпадение (10). В последних двух случаях перевод полностью неверный.

Предобработка предложений состоит в приведении к нижнему регистру и удалении знаков препинания, которые иначе могут быть оценены как совпадающие токены. Если для расчета метрики требуется предварительная токенизация, она производится с использованием метода split().

Таблица 1

Результаты оценки с использованием традиционных метрик

	Предложение	**BLEU (1,2,3,4)**	**Sacre BLEU**	**RIBES**	**TER**	**charac TER**	**chrf++**
1	Отец разбудил меня на утренней заре.	1.0, 1.0, 1.0, 1.0	100.0	1.0	0.0	0.0	100.0
2	На утренней заре меня разбудил отец.	1.0, 0.63, 0.47, 0.0	82.78	0.2	50.0	0.36	72.79
3	Папа разбудил меня на утренней заре.	0.83, 0.82, 0.79, 0.76	85.94	0.96	16.67	0.11	85.4
4	Мой отец разбудил меня на утренней заре.	0.86, 0.85, 0.83, 0.81	90.46	0.96	16.67	0.1	97.86
5	Отец разбудил меня утром на заре.	0.83, 0.58, 0.44, 0.0	66.04	0.57	33.33	0.31	61.06
6	Отец разбудил меня на заре.	0.82, 0.71, 0.65, 0.58	64.07	0.59	16.67	0.35	65.77
7	Отец не разбудил меня на утренней заре.	0.86, 0.76, 0.7, 0.64	88.28	0.96	16.67	0.08	89.85
8	Мать разбудила меня на утренней заре.	0.67, 0.63, 0.59, 0.51	78.39	0.9	33.33	0.14	75.89
9	У меня есть летний чум.	0.16, 0.0, 0.0, 0.0	9.44	0.0	83.33	1.0	12.31
10	У нас есть летний чум.	0, 0, 0, 0	3.4	0.0	100.0	1.0	8.54

BLEU рассчитывается на 4 уровнях (от униграмм до 4-грамм) с использованием библиотеки NLTK, из которой импортируется sentence_bleu, т. е. оценка на уровне предложения, а не текста. Как показывает таблица 1, получены оценки от максимальной для первого предложения до нуля для последнего. Самые высокие оценки получены на уровне униграмм, при этом перестановка слов вовсе не отразилась на оценке. Данная метрика оценивает пересечения n-грамм, поэтому она не может полностью отразить качество перевода. В частности, при добавлении отрицания (7) оценка оказалась выше, чем при изменениях, не повлиявших на смысл: удалении прилагательного, значение которого понятно из контекста (6), замене существительного на синоним (3).

Интересно, что для последнего предложения оценка SacreBLEU оказалась не равна нулю, хотя она и очень низкая. Вероятно, это связано с особенностями выравнивания, так как BLEU и SacreBLEU предназначены изначально для оценки целого текста, а не отдельных предложений, и SacreBLEU не имеет отдельной реализации для оценки предложений.

Что касается оценки метрики RIBES, она резко падает на предложении с измененным порядком слов. В данном случае для русского языка оно влияет на тема-рематическое членение предложения, но не меняет фактическую информацию. Оценка RIBES будет важна для направления перевода русский-ненецкий, так как ненецкий язык имеет фиксированный порядок слов.

Для TER и characTER, в отличие от остальных метрик, минимальное значение желательнее максимального (они рассчитывают количество вносимой в МП правки). Если привести значения TER и characTER к общему знаменателю и сопоставить их, видно, что оценки различаются между собой, так как TER ведет расчет по словам, а characTER – по символам.

Поскольку chrf++ учитывает присутствие единиц эталонного перевода в МП, наиболее высокую оценку (кроме полного совпадения) получает предложение (4), куда было добавлено притяжательное местоимение.

Последняя из рассматриваемых метрик – BERTScore.

Таблица 2

Результаты оценки с использованием BERTScore

	Предложение	**Precision**	**Recall**	**F1**
1	Отец разбудил меня на утренней заре.	1.0	1.0	1.0
2	На утренней заре меня разбудил отец.	0.9	0.9	0.9
3	Папа разбудил меня на утренней заре.	0.98	0.98	0.98
4	Мой отец разбудил меня на утренней заре.	0.94	0.98	0.96
5	Отец разбудил меня утром на заре.	0.94	0.9	0.92
6	Отец разбудил меня на заре.	0.97	0.88	0.92
7	Отец не разбудил меня на утренней заре.	0.97	0.99	0.98
8	Мать разбудила меня на утренней заре.	0.97	0.98	0.98
9	У меня есть летний чум.	0.71	0.67	0.69
10	У нас есть летний чум.	0.71	0.67	0.69

Интересно, что значение метрики BERTScore ни для одного из предложений не равно нулю, хотя последнее предложение не имеет ничего общего с эталонным. Также обращает на себя внимание тот факт, что высокую оценку получили предложения 7 и 8, содержащие фактические ошибки.

Для оценки МП в направлении «русский-ненецкий» BERTScore использоваться не может, так как необходима поддержка языка перевода.

Использование набора метрик позволяет не только оценить МП с разных позиций, но и привлечь для анализа оценки, полученные для языковых пар, включающих другие языки коренных народов России. Например, для оценки МП в языковой паре «татарский-русский» для оценки результатов используется BLEU [22], а в языковой паре «эрзянский-русский» – BLEU и ChrF++ [23].

Отметим, что время расчета для метрик различается. Поскольку время может зависеть от таких факторов, как загруженность системы, код был запущен 5 раз для расчета среднего показателя. Показатели могут измениться, если код будет запускаться более 10 раз и на разных графических процессорах.

Таблица 3

Время расчета метрики для 10 предложений

Метрика	**Время расчета, сек.**					**Среднее время, сек**
BLEU	0,005	0,010	0,006	0,007	0,005	0,006
SacreBLEU	0,077	0,090	0,040	0,063	0,004	0,054
RIBES	0,008	0,004	0,004	0,005	0,004	0,005
TER	0,045	0,039	0,065	0,058	0,041	0,005
characTER	0,078	0,032	0,033	0,037	0,050	0,046
chrf++	0,156	0,039	0,036	0,053	0,038	0,064
BERTScore	3,846	25,310	2,664	4,532	2,359	7,742

Как показывает таблица 3, чем больше параметров учитывает метрика, тем больше времени требуется на вычисление. Нейросетевая метрика BERTScore требует для расчета результатов гораздо больше времени и ресурсов, по сравнению с использованием классических метрик.

Представленный набор из 10 предложений моделирует МП, позволяя определить, как на оценки метрик влияют разные отклонения от эталонного перевода (варианты и ошибки). В дальнейшем для оценки результатов МП используется набор реальных предложений, представляющих разные типы текста в наборе данных [24]. Предложения, включенные в этот контрольный набор, удаляются из набора данных перед обучением модели МП.

Использование метрик автоматической оценки качества МП позволяет быстро получить непредвзятую оценку для контрольного набора предложений. Целесообразность использования набора метрик определяется направлением перевода для малоресурсных языков. Кроме того, использование метрик «золотого стандарта» позволяет вписать исследование в систему публикаций связанной тематики.

Особенность исследования малоресурсных языков состоит в том, что для оценки качества МП могут использоваться только классические метрики, которые не требуют поддержки конкретного языка. Разработка нейросетевых метрик для малоресурсных языков остается делом ближайшего будущего, поскольку на настоящем этапе для многих из них стоит задача сбора данных. Параллельно будет происходить совершенствование систем МП.

ЛИТЕРАТУРА

1. Babych B. Automated MT evaluation metrics and their limitations [Electronic resource] // Revista Tradumàtica: tecnologies de la traducció. 2014. № 12, Traducció i qualitat. P. 464–470. URL: https://www.raco.cat/index.php/Tradumatica/article/download/286868/375090 (date of access: 14.07.2025).
2. Marie B., Fujita A., Rubino R. Scientific Credibility of Machine Translation Research: A Meta-Evaluation of 769 Papers [Electronic resource] // Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing. 2021. P. 7297–7306. DOI: https://doi.org/10.18653/v1/2021.acl-long.566.
3. The NIST 2008 Metrics for machine translation challenge–overview, methodology, metrics, and results [Electronic resource] / M. Przybocki, K. Peterson, S. Bronsart, G. Sanders // Machine Translation. 2009. № 23(2/3). P. 71–103. DOI: https://doi.org/10.1007/s10590-009-9065-6.
4. Нуриев В. А., Егорова А. Ю. Методы оценки качества машинного перевода: современное состояние // Информатика и ее применение. 2021. Том 15. Выпуск 2. С. 104–111.
5. Берендяев М. В., Гилин М. И., Коканова Е. С. Генеративный искусственный интеллект и оценка качества перевода // Человек: Образ и сущность. 2025. №2(62). [в печати]
6. BLEU: a method for automatic evaluation of machine translation [Electronic resource] / K. Papineni, S. Roukos, T. Ward, W.-J. Zhu // Proceedings of the 40th Annual Meeting on Association for Computational Linguistics (ACL '02). Association for Computational Linguistics, USA, 2002. P. 311–318. DOI: https://doi.org/10.3115/1073083.1073135.
7. Коканова Е. С., Поротова П. В. Автоматическая метрика оценка качества машинного перевода BLEU // Профессия переводчика: вызовы и перспективы: Сборник студенческих научных статей по материалам работы III Всероссийской молодежной научной онлайн-конференции «Современные техно-

логии в переводе. Машинный перевод: от стереотипов к новым возможностям», Архангельск, 24 ноября 2022 года. Архангельск: Северный (Арктический) федеральный университет им. М. В. Ломоносова, 2023. С. 44–47.
8. Post M. A Call for Clarity in Reporting BLEU Scores [Electronic resource] // Proceedings of the Third Conference on Machine Translation: Research Papers. 2018. P. 186–191. DOI: https://doi.org/10.18653/v1/W18-6319.
9. Doddington G. Automatic Evaluation of Machine Translation Quality Using N-gram Co-Occurrence Statistics [Electronic resource] // HLT '02: Proceedings of the second international conference on Human Language Technology Research. 2002. P. 138–145. URL: https://dl.acm.org/doi/10.5555/1289189.1289273 (date of access: 14.07.2025).
10. Banerjee S., Lavie A. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments [Electronic resource] // Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization. 2005. P. 65–72. URL: https://aclanthology.org/W05-0909/ (date of access: 14.07.2025).
11. WordNet [Electronic resource]. URL: https://wordnet.princeton.edu/ (date of access: 14.07.2025).
12. Automatic Evaluation of Translation Quality for Distant Language Pairs [Electronic resource] / H. Isozaki, T. Hirao, K. Duh, K. Sudoh, H. Tsukada // Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing. 2010. P. 944–952. URL: https://aclanthology.org/D10-1092/ (date of access: 14.07.2025).
13. Lin C.-Y. ROUGE: A Package for Automatic Evaluation of Summaries [Electronic resource] // Annual Meeting of the Association for Computational Linguistics. 2004. URL: https://aclanthology.org/W04-1013/ (date of access: 14.07.2025).
14. Study of Translation Edit Rate with Targeted Human Annotation [Electronic resource] / M. Snover, B. Dorr, R. Schwartz, L. Micciulla, J. Makhoul // Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers. 2006. P. 223–231. URL: https://aclanthology.org/2006.amta-papers.25/ (date of access: 14.07.2025).
15. CharacTer: Translation Edit Rate on Character Level [Electronic resource] / W. Wang, J.-T. Peter, H. Rosendahl, H. Ney // Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers. 2016. P. 505–510. DOI: https://doi.org/10.18653/v1/W16-2342.
16. Popović M. CharacTer: chrF++: words helping character n-grams [Electronic resource] // Proceedings of the Second Conference on Machine Translation. 2016. P. 612–618. DOI: https://doi.org/10.18653/v1/W17-4770.
17. Results of WMT22 Metrics Shared Task: Stop Using BLEU – Neural Metrics Are Better and More Robust [Electronic resource] / M. Freitag, R. Rei, N. Mathur, C.-k. Lo, C. Stewart, E. Avramidis, T. Kocmi, G. Foster, A. Lavie, A. F. T. Martins // Proceedings of the Seventh Conference on Machine Translation (WMT). 2022. P. 46–68. URL: https://aclanthology.org/2022.wmt-1.2/ (date of access: 14.07.2025).

18. COMET: A Neural Framework for MT Evaluation [Electronic resource] / R. Rei, C. Stewart, A. C. Farinha, A. Lavie // Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2020. P. 2685–2702. DOI: https://doi.org/10.18653/v1/2020.emnlp-main.213.
19. BERTScore: Evaluating Text Generation with BERT [Electronic resource] / T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, Y. Artzi. 2020. DOI: https://doi.org/10.48550/arXiv.1904.09675.
20. The Inside Story: Towards Better Understanding of Machine Translation Neural Evaluation Metrics [Electronic resource] / R. Rei, N. M. Guerreiro, M. Treviso, L. Coheur, A. Lavie, A. F. T. Martins // Proceedings of the Seventh Conference on Machine Translation (WMT). 2023. P. 46–68. DOI: DOI: https://doi.org/10.18653/v1/2023.acl-short.94
21. Кипяткова И. С., Кагиров И. А., Долгушин М. Д. Применение предварительно обученных многоязычных моделей для распознавания карельской речи [Электронный ресурс] // Информатика и автоматизация. 2025. № 24 (2). С. 604-630. DOI: https://doi.org/10.15622/ia.24.2.9.
22. Application of Low-resource Machine Translation Techniques to Russian-Tatar Language Pair [Electronic resource] / A. Valeev, I. Gibadullin, A. Khusainova, A. Khan. 2019. DOI: https://doi.org/10.48550/arXiv.1910.00368.
23. Dale D. The first neural machine translation system for the Erzya language [Electronic resource] // Proceedings of the first workshop on NLP applications to field linguistics. 2022. P. 45–53. URL: https://aclanthology.org/2022.fieldmatters-1.6/ (date of access: 14.07.2025).
24. Епимахова А. С., Коканова Е. С. Тундровый ненецкий язык: несбалансированность набора данных // СЛОВА и ЦИФРЫ: материалы I Международной научно-практической конференции. Йошкар-Ола – Москва, 4–5 апреля 2025 г. / Министерство науки и высшего образования Российской Федерации, Поволжский государственный технологический университет, Университет науки и технологий МИСИС. Йошкар-Ола: Поволжский государственный технологический университет, 2025. С. 110–115.