УДК 81’322 DOI: 10.31249/chel/2025.02.10

Берендяев М.В.[1)], **Гилин М.И.**[2)], **Коканова Е.С.**[1)]

ГЕНЕРАТИВНЫЙ ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ И ОЦЕНКА КАЧЕСТВА ПЕРЕВОДА©

[1)] *Северный* (*Арктический*) *федеральный университет имени М.В. Ломоносова,*
Россия, *Архангельск*, *m.berendyaev@narfu.ru, e.s.kokanova@narfu.ru,*
[2)] *Университет науки и технологий МИСИС*,
Россия, *Москва*, *m.gilin@misis.ru,*

Аннотация. В статье рассматриваются подходы к оценке качества перевода в контексте использования автоматической генерации текста и, в частности, машинного перевода. Предложена новая метрика оценки качества автоматически сгенерированного текста, исходя из прогнозируемого времени его постобработки, в основе которой лежит оценка трудоемкости исправления ошибок и рисков влияния ошибок на достижение целей переводческого проекта. Эта метрика призвана решить проблемы оценки качества автоматического перевода, появившиеся с приходом генеративного искусственного интеллекта на смену машинному переводу.

Ключевые слова: оценка качества перевода; машинный перевод; автоматическая генерация текста; метрика; прогнозируемое время постобработки; генеративный искусственный интеллект.

Получена: 15.07.2024 Принята к печати: 28.12.2024

Berendyaev M.V.[1)], **Gilin M.I.**[2)], **Kokanova E.S.**[1)]

Generative AI and evaluation of translation output©

[1)] *Northern* (*Arctic*) *Federal University*, *Russia*, *Arkhangelsk*,
m.berendyaev@narfu.ru, *e.s.kokanova@narfu.ru*
[2)] *MISIS University of Science and Technology*,
Russia, *Moscow*, *m.gilin@misis.ru*

Abstract. The paper reviews approaches to evaluation of translation output in the context of using machine translation and automatic text generation systems. A new metric for assessing the quality of automatically generated text based on the predicted distance of its post-processing is proposed. The metric is built around the labor intensity of error correction and the risks of error impact on achieving the goals of a translation project. This metric aims to address the problems of evaluating the quality of automatic translation that have emerged with the advent of generative artificial intelligence replacing machine translation.

Keywords: evaluation of translation output; translation quality assessment; machine translation; automatic text generation; metric; predicted post-process time; generative artificial intelligence.

Received: 15.07.2024 Accepted: 28.12.2024

Введение

Оценка качества перевода вообще и качества выдачи систем автоматического перевода в частности – сложный процесс, который вызывает большой интерес у исследователей и практиков. В цифровую эпоху при его описании нельзя обойтись без определения базовых понятий, к которым относятся автоматическая генерация текстов (АГТ), большая языковая модель (БЯМ), постобработка результата АГТ. Под автоматической генерацией текстов в профессиональном переводе сегодня понимается создание текстов компьютерными программами в автоматическом режиме без участия или при минимальном участии человека [Берендяев, Светова, Коканова, 2024а]. А в последних итерациях отраслевых стандартов по переводу предпринимаются попытки заменить понятие «текст» на понятие *translation output* («выдача») и *target translation content* [ISO 5060:2024, p. 2] («контент») и рассматривать переводческий и лока-

лизационный проект в комплексе. К переводу текста с одного естественного языка на другой компьютерными программами без вмешательства человека относятся не только классический машинный перевод (МП), но и результаты выдачи генеративной нейросети. В русскоязычной литературе предлагаем использовать термин «большие языковые модели» (сокр. БЯМ) как эквивалент для *large language models* (сокр. LLM) [Меморандум …], под которым понимается комплексная нейромодель, использующая искусственные нейронные сети на основе трансформеров, обученная на значительных объемах текстовых данных методом самостоятельного обучения с модулем интерактивного диалога с пользователем для генерации ответа на запрос на естественном языке [Language models are …, 2019; Lankford, Way, 2024]. Под постобработкой результата АГТ понимается процесс доведения результата выдачи систем АГТ до требуемого уровня качества [Берендяев, Светова, Коканова, 2024б].

С одной стороны, стоимость использования генеративных нейросетей, будучи на порядки меньше, нежели систем МП, привлекает заказчиков перевода и переводческие предприятия, которые стремятся сократить расходы на постобработку результата АГТ с участием человека без ущерба качеству итогового продукта [Берендяев, Светова, Коканова, 2024а]. С другой стороны, время на выдачу готового к использованию перевода нередко увеличивается в десятки и сотни раз. В настоящий момент генеративные нейросети не всегда пригодны для перевода в режиме реального времени [How much data …, 2024]. И сам автоматически сгенерированный текст может оказаться непредсказуемого качества [The State of Machine Translation, 2024]. Следует также отметить, что на наших глазах меняются стратегии постобработки результата АГТ и сценарии работы с автоматически сгенерированными текстами с целью наиболее эффективного доведения их до требуемого уровня качества. А с приходом в профессию нового поколения специалистов-выпускников подходы, ранее использовавшиеся для редактирования перевода, выполненного человеком, все меньше применяются для работы с результатами МП и АГТ. Цель работы участников исследования, как и многих отраслевых объединений, – оставить в прошлом этот метод и перестать работать с нейросетями по тем же правилам, что и с ручным переводом, выполненным человеком.

Метрики оценки качества перевода

Не ставя перед собой цели дать аналитический обзор всех существующих метрик качества перевода, что само по себе потребовало бы отдельного исследования, авторы приняли решение ограничиться некоторыми принципиальными моментами.

Во-первых, англоязычный термин *assessment* часто используется как синоним термина *evaluation*. Последний более предпочтителен, так как именно «оценка качества перевода» ориентирована на продукт и критический анализ оценки. Термин «анализ ошибок в тексте перевода» ориентирован на процесс и описание типологии ошибок. Предпочтение англоязычному термину *evaluation* также отдается потому, что он помогает избежать путаницы, которую вызывает аббревиатура QA, от *quality assurance* («проверка и обеспечение качества перевода») [ISO 5060:2024, p. 2].

Во-вторых, на момент подготовки публикации не существует консенсуса и единой стандартизации в отношении метрик оценки качества перевода. Оценку качества перевода, выполненного человеком, и экспертную оценку качества МП (например, шкала Ликерта, HTER (Human Translation Edit Rate), а также DQF / MQM (Dynamic Quality Framework / Multidimensional Quality Metrics)) нельзя считать применимыми для оценки качества результата выдачи АГТ. Автоматические метрики (WER, TER, BLUE и другие) также не сумели показать степень релевантности и адекватности оценки качества выдачи АГТ.

Автоматическая оценка качества МП измеряет уровень соответствия МП эталонному, или референсному переводу (РП). Для определения уровня соответствия МП и РП применяются критерии точности и полноты [Нуриев, Егорова, 2021]. Одна из первых метрик, Word Error Rate (WER), была основана на расстоянии Левенштейна (расстоянии редактирования, *edit distance*). Количество итераций редактирования делилось между количеством слов в РП. Таким образом, расстояние Левенштейна – это метрика сходства между двумя строковыми последовательностями [Kompe, 1997, p. 27]. Например, расстояние между «Австрия» и «Австралия» по Левенштейну составит два, так как понадобится выполнить два удаления. Основным недостатком такой метрики являлось то, что она не учитывает перестановку слов и замены, а удаления и вставки будут равны [IDIAP Research Report …, 2005, p. 2–3].

Еще одной получившей широкое распространение метрикой автоматической оценки качества МП стала TER (Translation Error Rate). Она исходит из расчета исправлений, необходимых для приведения МП к РП. При этом пунктуационные знаки принимаются за отдельные слова, а исправлениями считаются не только удаление, вставка и замена, но и перестановка [Agarwal, Lavie, 2008, p. 115–116].

BLEU (Bilingual Evaluation Understudy) также не учитывает семантическую близость и контекст, а фокусируется только на n-граммном совпадении. Если при переводе передан тот же смысл, но присутствует значительное несовпадение n-грамм в МП и РП, качество перевода будет оценено низко [Митренина, 2017, с. 184]. Из вышеизложенного следует, что BLEU глубоко зависима от качества и стиля РП. При условии, что эталонный текст перевода не является единственно верным, оценка может не отражать действительную качественную разницу. К основным минусам метрики можно отнести недостаточный учет семантики и контекста, чрезмерную зависимость от качества референсного перевода и невозможность оценить его адекватность. BLEU может быть использована для быстрого сравнения эффективности применения разных систем МП. На наш взгляд, для более комплексного оценивания качества выдачи систем МП и АГТ видится предпочтительным ее использование в сочетании с другими автоматическими метриками (например, BERT, COMET, UniTE, а также безреференсными *reference-free metrics*) и экспертной оценкой, что также подразумевает возрастание временных затрат и когнитивных усилий в случае большого объема данных к переводу.

В-третьих, теоретически перспективным оказывается вопрос о подходах к оценке качества перевода. Принято выделять два подхода: холистический и аналитический. При первом основное внимание уделяется анализу качества всего перевода, а не разбору ошибок в каждом отдельном сегменте. Второй подход предполагает сравнение содержания целевого языка с содержанием исходного языка на основе сегментов с учетом специфики переводческого проекта. Несоответствия рассматриваются как ошибки. Штрафные баллы присваиваются каждой ошибке (выделяется семь основных видов и 34 подвида), а также учитывается степень критичности ошибки в переводе [ISO 5060:2024, p. 7].

Таким образом, на данном этапе рассуждений можно говорить о релевантности следующих утверждений.

- Использование общих метрик недопустимо. Применимость метрики и оценка качества перевода на одном тексте (фрагменте) не означает ее применимости и тождественности оценке в рамках проекта, состоящего из множества текстов (фрагментов).
- Метрики должны разрабатываться индивидуально для каждого типа текста (контента), языкового направления, объема.
- Существующие метрики оценки качества перевода, выполненного человеком, выдачи систем МП и АГТ, а также результата их сочетания, бесполезны при работе с БЯМ. В связи с этим авторы статьи предлагают взять за основу холистический подход при оценке качества автоматически сгенерированного текста.

Авторы полагают, что оценка качества выдачи систем МП и АГТ на основе сравнения с референсным переводом не имеет практического смысла в силу самой природы перевода и сосуществования множества верных переводов для любого исходного текста нетривиальной длины (для одного токена / сегмента / слова / семантической единицы может быть сформировано несколько верных переводов в зависимости от требований заказчика, коммуникативной ситуации дискурса и т.д.). Кроме того, оценка на основе *edit distance* излишне механизирована, а оценка качества выдачи систем МП и АГТ на основе подсчета ошибок «загрязнена» субъективностью и потенциальной усталостью эксперта, отсутствием четкой классификации (типологии) и гарантий верного отнесения к той или иной ошибке, а также пересечением ошибок. Агрегирование оценки на основе сегмента / предложения / сверхфразового единства / фрагмента с ее экстраполяцией на объем всего текста или проекта не позволяет получать воспроизводимые результаты. Авторы видят это ненадежным и крайне субъективным.

Вышеизложенное позволяет сделать важное обобщение. Оценка качества выдачи систем МП и АГТ должна разрабатываться индивидуально:

- для каждого языкового направления с учетом локали (например, направление английский – русский не равно направлению русский – английский);
- для каждой предметной области / заказчика / проекта и т.д.;
- для каждого типа текста (контента);
- для каждого типового технологического процесса.

На наш взгляд, чрезвычайный разнобой в подходах к оценке качества перевода в цифровую эпоху обусловлен в первую очередь нерешенностью вопроса о разграничении упомянутых выше понятий «текст» и «контент», а также отсутствием у заказчика переводческого или локализационного проекта возможности четко, понятно и окончательно сформулировать свои требования к качеству на фоне общей тенденции к экономии ресурсов [Leveraging GPT-4 …, 2023]. Авторы считают, что лишь определив текст как часть переводческого проекта, можно надеяться на успех в анализе оценки качества результата выдачи МП и АГТ. Например, с помощью предлагаемой метрики «прогнозируемое время постобработки автоматически сгенерированного текста», основанной на трудоемкости.

Разработка метрики

Необходимо подчеркнуть, что оценка качества выдачи систем АГТ классическими методами посредством формул качества, основанных на выявлении лингвистических (языковых), технических, типографических и дизайнерских ошибок, не может считаться достаточно эффективной, поскольку в таком случае текст (контент) оценивается как готовый продукт, который должен конкурировать с таким же готовым переводом, выполненным человеком. Однако авторы уверены в некорректности такого подхода и настаивают на необходимости оценивать системы АГТ по такому же принципу, по которому оцениваются другие фасилитирующие работу инструменты (например, интерфейсы CAT, таблицы Excel, макросы в MS Word), то есть рассматривать системы АГТ в первую и главную очередь как инструмент, позволяющий быстрее получить готовый продукт на выходе, а не инструмент, генерирующий готовый продукт. Другими словами, оценивать, насколько быстрее и дешевле инструмент позволяет выполнять поставленную задачу (в случае с лингвистическими услугами – перевод), а не насколько грамотно и (или) точно система выполняет перевод текста.

Например, при переводе с русского языка на английский текста «Были составлены подробные инструкции для ответственных лиц» в выдаче системы автоматического перевода Google NMT *Detailed guidance have been produced for accountants* содержится грубая грамматическая ошибка, которая при использовании в оценке перевода классических формул качества (например, экспертной) неминуемо существенно снизила бы итоговый балл такому переводу.

К подобным ошибкам можно причислить ошибки в именах, простых названиях и цифрах. А если система АГТ выдает перевод *Placed tablets are mostly 11 pieces in one blister* для исходного текста «Размещаются таблетки в основном по 10 штук в одном блистере», то в случае оценки общепринятыми формулами качества итоговый балл такого перевода будет крайне низким. Однако такая ошибка не вызывает трудностей в распознавании, а ее исправление занимает несколько секунд.

В то же время пунктуационные и терминологические ошибки оцениваются как менее критичные, однако могут требовать значительного времени на исправление в случае сложносочиненных предложений в сегментах для первого случая и узкоспециализированной терминологии – для второго. Например, предложение *We, as a Council, decided that UNMIL required specific assets to fulfil its mandate, and we must honour our commitment to ensuring that it has the tools on hand to do its job* было переведено системой АГТ как «Мы, члены Совета, постановили, что МООНЛ требуются специальные активы для выполнения своего мандата и мы должны выполнить свое обязательство и обеспечить ей все необходимые средства для выполнения своих задач». Как видно из примера, в предложении отсутствует запятая после фразы «для выполнения своего мандата», однако нахождение этой ошибки в силу сложносочиненной природы предложения может быть затруднено как необходимостью перечитывать оба сегмента (исходного текста и перевода) несколько раз, так и вероятностью пропуска этой ошибки при первичной постобработке, что потенциально еще больше увеличит абсолютное время на редактирование сегмента. В случае отсутствия перевода аббревиатуры UNML в глоссарии проекта итоговое время на редактирование сегмента увеличится еще и за счет необходимости перепроверить правильность перевода термина вне зависимости от его корректности или некорректности в тексте перевода, что также не учитывается формулами качества и не может учитываться даже в теории.

За единицу измерения прогнозируемого времени постобработки был принят «условный временной промежуток» – время, необходимое специалисту по постобработке автоматически сгенерированных текстов / редактору на исправление самой незначительной с точки зрения трудоемкости исправления ошибки. Таким образом, авторы стремились снизить вероятность искажения результатов оценки человеческим фактором, заключавшимся в раз-

личающихся навыках лингвистов использовать вычислительную технику и сеть Интернет. Предполагается, что время, требуемое на работу с компьютером по набору текста на клавиатуре, поиску терминологии в Интернете и в целом скорость выполнения работы могут сильно разниться между исполнителями, однако уровень перечисленных навыков для одного человека одинаково влияет на скорость выполнения работы для отдельно взятого человека.

В зависимости от группы ошибок количество условных временных промежутков (УВП) для исправления каждой группы будет разниться и составит 1–3 УВП для «короткой» ошибки, 4–6 УВП для «средней» и 7–9 для «длинной». Такие значения были приняты эмпирически после выполнения оценки тестовых массивов.

Таким образом, формула прогнозируемого времени постобработки автоматически сгенерированного текста в базовом варианте выглядит следующим образом:

PPT = Nshort*short + Nmidterm*midterm + Nlong*long,

где PPT – прогнозируемое время постобработки (*predicted post-process time*), Nn – количество ошибок класса, short – коэффициент трудоемкости исправления «коротких» ошибок, midterm – коэффициент трудоемкости исправления «средних» ошибок и long – коэффициент трудоемкости исправления «длинных» ошибок.

Таблица

Опасность риска

Ошибки	Коэффициент риска для проекта / Трудоемкость исправления	0	1	2	3	4	5	6	7	8	9	10
Короткая ошибка	1 УВП	0	1	2	3	4	5	6	7	8	9	10
	2 УВП	0	2	4	6	8	10	12	14	16	18	20
	3 УВП	0	3	6	9	12	15	18	21	24	27	30
Средняя ошибка	4 УВП	0	4	8	12	16	20	24	28	32	36	40
	5 УВП	0	5	10	15	20	25	30	35	40	45	50
	6 УВП	0	6	12	18	24	30	36	42	48	54	60
Длинная ошибка	7 УВП	0	7	14	21	28	35	42	49	56	63	70
	8 УВП	0	8	16	24	32	40	48	56	64	72	80
	9 УВП	0	9	18	27	36	45	54	63	72	81	90

При нынешнем потоке оперативной и быстро устаревающей текстовой информации (комплексного контента) на фоне сетевизации всех сфер деятельности человека у профессиональных участников рынка перевода и локализации не всегда есть время и ресурсы последовательно и полностью обрабатывать, редактировать и исправлять весь текст (контент). В связи с этим, когда речь идет о процессах, сопряженных с рисками, как и практически в любой сфере профессиональной деятельности, при постобработке автоматически сгенерированного текста в профессиональном переводе следует также вести обработку (выявление и исправление) от ошибок с наиболее высоким уровнем риска к ошибкам с меньшими рисками для прагматических целей того проекта, в интересах которого осуществляется перевод. Линейная шкала в таблице демонстрирует, что одна и та же «длинная» ошибка может иметь значение «риск*трудоемкость» от 0 до 90, поэтому в разработанную метрику добавлено дополнительное измерение – риск для проекта, то есть опасность ошибки для целей заказчика.

В рамках представленной системы оценки очевидно, что исправление десяти «коротких» ошибок с высоким риском более полезно для проекта, чем одной «длинной» ошибки с низким риском. Авторы считают, что цель практикующего специалиста по постобработке автоматически сгенерированного текста должна быть той же, что и у профессионального технолога перевода – не упустить важного и не добавить лишнего, правильно спрогнозировать объем работ, рассчитать силы и время и выполнить проект в указанный срок.

Заключение

В настоящий момент становится все более острой необходимость новых подходов к оценке качества выдачи систем МП и АГТ, что заставляет разработчиков современных стандартов в области АГТ четче понимать, с какой целью измерять качество:

- для выбора и сравнения систем АГТ (в том числе МП) между собой;
- для сравнения конкретных переводчиков, редакторов, специалистов по постобработке автоматически сгенерированных текстов, переводческих предприятий между собой;
- для выявления пользы или вреда от обучения систем МП и АГТ;

- для принятия корректирующих мер и обратной связи, претензий переводчикам, редакторам, переводческим компаниям;
- для дидактических целей и т.д.

Чем сложнее и обширнее данные, на которых обучаются нейросетевые модели, тем менее предсказуемы результаты их дальнейшей работы. В свете стремительного развития технологий генеративного искусственного интеллекта (ИИ) попытки участников рынка лингвистических услуг сохранить баланс между качеством и итоговой стоимостью перевода должны быть более оптимизированными и генерализованными. В этих условиях создание индивидуальных метрик оценки трудоемкости и качества выдачи систем МП и АГТ в профессиональном переводе до осуществления собственно постобработки автоматически сгенерированных текстов, равно как и оценка успешности такой постобработки, могут быть в будущем переданы современным высокопроизводительным инструментам – нейросетевым языковым моделям и генеративному ИИ. В связи с этим необходима совместная с участниками рынка разработка правил, принципов и запросов (промптов), которые позволяли бы поручать ИИ работу по формированию метрик оценки качества in vivo для конкретного проекта и текста (контента), коммуникативной ситуации и уровня обученности других систем.

Неясной перспективой остается делегирование ИИ задач по оценке уместности изменений, исправлений, редактирования результата АГТ в ходе постобработки автоматически сгенерированного текста (контента), а также ответ на ключевой вопрос: изменился ли смысл после обработки автоматически сгенерированного текста (контента) и были ли оправданны временные и трудовые затраты на выполнение такой работы.

Разработанная метрика «прогнозируемое время постобработки автоматически сгенерированного текста» – это проектоориентированная метрика оценки качества выдачи систем МП и АГТ, которая учитывает не «стоимость» ошибки, а время, которое может занять ее исправление. В качестве продолжения исследования видится добавление третьего измерения вероятности возникновения ошибки помимо длины исправления ошибки и риска, поскольку вероятность возникновения ошибки в автоматически сгенерированном тексте может разительно отличаться от вероятности ее возникновения в переводе, выполненном человеком.

Список литературы

Берендяев М.В., *Светова С.Ю.*, *Коканова Е.С.* Автоматическая генерация текстов в профессиональном переводе // Didactica Translatorica. – 2024а. – № 2. – С. 5–10. – URL: https://didact-translat.ru/upload/iblock/39f/tml2l09pcmrtl5qprtpjk0kfrt2eslp9.pdf

Берендяев М.В., *Светова С.Ю.*, *Коканова Е.С.* Постобработка результатов автоматической генерации перевода // Развитие Севера и Арктики: формирование и сохранение традиционных российских духовно-нравственных ценностей / отв. ред. Л.Ю. Щипицина. – Архангельск : САФУ, 2024б. – С. 182–185.

Меморандум Ассоциации переводческих компаний по вопросам институционализации, стандартизации и создания нормативно-правовой базы для применения машинного перевода и прочих технологий автоматической генерации текста в качестве профессиональных инструментов в области перевода и локализации // Ассоциация переводческих компаний. – URL: https://atcru.org/upload/iblock/ff5/fp58d29psjlsr3minttu0nuqqntlay50.pdf

Митренина О.В. Машинный перевод // Прикладная и компьютерная лингвистика. – Москва : Ленанд, 2017. – С. 156–189.

Нуриев В.А., *Егорова А.Ю.* Методы оценки качества машинного перевода: современное состояние // Информатика и ее применение. – 2021. – Т. 15, Вып. 2. – С. 104–111. DOI: https://doi.org/10.14357/19922264210215

Agarwal A., *Lavie A.* Meteor, M-BLEU and M-TER: evaluation metrics for high-correlation with human rankings of machine translation output // Proceedings of the Third Workshop on Statistical Machine Translation. – Columbus : Ohio, 2008. – P. 115–118. – URL: https://aclanthology.org/W08-0312.pdf

How much data is enough data? Fine-tuning large language models for in-house translation: performance evaluation across multiple dataset sizes. Vieira I., Lankford W.A.S., Castilho S., Way A. // AMTA. – 2024. – Vol. 1. – P. 236–249.

IDIAP Research Report 04–73. On the use of information retrieval measures for speech recognition evaluation / McCowan I., Moore D., Dines J., Gatica-Perez D., Flynn M., Wellner P., Bourlard H. – 2005. – URL: https://publications.idiap.ch/downloads/reports/2004/rr04-73.pdf

ISO 5060:2024. Translation services – Evaluation of translation output – General guidance. – URL: https://www.iso.org/obp/ui/en/#iso:std:iso:5060:ed-1:v1:en

Kompe R. Prosody in speech understanding systems. – Berlin ; Heidelberg ; New York ; Barcelona ; Budapest ; Hong Kong ; London ; Milan ; Paris ; Santa Clara ; Singapore ; Tokyo : Springer, 1997. – 359 p.

Language models are unsupervised multitask learners // Radford A., Wu J., Child R., Luan D., Amodei D., Sutskever I. – 2019. – URL: https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf

Lankford S., *Way A*. Leveraging LLMs for MT in crisis scenarios: a blueprint for low-resource languages // AMTA. – 2024. – Vol. 1. – P. 4–13.

Leveraging GPT-4 for automatic translation post-editing / Raunak V., Sharaf A., Wang Y., Awadalla H., Menezes A. // Findings of the Association for Computational Linguistics: EMNLP 2023. – 2023. – URL: https://arxiv.org/abs/2305.14878

The state of machine translation 2024. – URL: https://inten.to/machine-translation-report-2024/

References

Berendjaev, M.V., Svetova, S.Ju., Kokanova, E.S. (2024b). Postobrabotka rezul'tatov avtomatichekoj generatsii perevoda. In *Razvitie Severa i Arktiki*: *formirovanie i sohranenie traditsionnyh rossijskih duhovno-nravstvennyh tsennostej* (pp. 182–185). Arhangel'sk: SAFU.

Berendjaev, M.V., Svetova, S.Ju., Kokanova, E.S. (2024a). Avtomaticheskaja generacija tekstov v professional'nom perevode. *Didactica Translatorica*, *2*, 5–10. Retrieved from: https://didact-translat.ru/upload/iblock/39f/tml2l09pcmrtl5qprtpjk0kfrt2eslp9.pdf

Memorandum Assotsiatsii perevodcheskih kompanij po voprosam institutsionalizatsii, standartizatsii i sozdanija normativno-pravovoj bazy dlja primenenija mashinnogo perevoda i prochih tehnologij avtomaticheskoj generatsii teksta v kachestve professional'nyh instrumentov v oblasti perevoda i lokalizatsii (2024). Retrieved from: https://atcru.org/upload/iblock/ff5/fp58d29psjlsr3minttu0nuqqntlay50.pdf

Mitrenina, O.V. (2017). Mashinnyj perevod. In *Prikladnaja i komp'uternaja lingvistika* (pp. 156–189). Moscow: Lenand.

Nuriev, V.A., Egorova, A.Ju. (2021). Metody otsenki kachestva mashinnogo perevoda: sovremennoe sostojanie. *Informatika i ejo primenenie*, *15*(*2*), 104–111. DOI: https://doi.org/10.14357/19922264210215

Agarwal, A., Lavie, A. (2008). Meteor, M-BLEU and M-TER: evaluation metrics for high-correlation with human rankings of machine translation output. In *Proceedings of the Third Workshop on Statistical Machine Translation* (pp. 115–118). Columbus: Ohio.

ISO 5060:2024. Translation services – Evaluation of translation output – General guidance. Retrieved from: https://www.iso.org/obp/ui/en/#iso:std:iso:5060:ed-1:v1:en

Kompe, R. (1997). *Prosody in speech understanding systems*. Berlin; Heidelberg; New York; Barcelona; Budapest; Hong Kong; London; Milan; Paris; Santa Clara; Singapore; Tokyo: Springer.

Lankford, S., Way, A. (2024). Leveraging LLMs for MT in crisis scenarios: a blueprint for low-resource languages. *AMTA*, *1*, 4–13.

McCowan, I., Moore, D., Dines, J., Gatica-Perez, D., Flynn, M., Wellner, P., Bourlard, H. (2005). IDIAP Research Report 04-73. On the use of information retrieval measures for speech recognition evaluation. Retrieved from: https://publications.idiap.ch/downloads/reports/2004/rr04-73.pdf

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I. (2019). Language models are unsupervised multitask learners. Retrieved from: https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf

Raunak, V., Sharaf, A., Wang, Y., Awadalla, H., Menezes, A. (2023). Leveraging GPT-4 for automatic translation post-editing. *Findings of the Association for Computational Linguistics*: *EMNLP 2023*. Retrieved from: https://arxiv.org/abs/2305.14878

The state of machine translation 2024. Retrieved from: https://inten.to/machine-translation-report-2024/

Vieira, I., Lankford, W.A.S., Castilho, S., Way, A. (2024). How much data is enough data? Fine-tuning large language models for in-house translation: performance evaluation across multiple dataset sizes. *AMTA*, *1*, 236–249.

Об авторах

Берендяев Максим Викторович – доцент, базовая кафедра технологий и автоматизации перевода в бюро переводов «АКМ-Вест», Северный (Арктический) федеральный университет имени М.В. Ломоносова, Россия, Архангельск, m.berendyaev@narfu.ru

Гилин Михаил Игоревич – ассистент, кафедра иностранных языков и коммуникативных технологий Института базового образования, Университет науки и технологий МИСИС, Россия, Москва, m.gilin@misis.ru

Коканова Елена Сергеевна – кандидат филологических наук, доцент, заведующий базовой кафедрой технологий и автоматизации перевода в бюро переводов «АКМ-Вест», Северный (Арктический) федеральный университет имени М.В. Ломоносова, Россия, Архангельск, e.s.kokanova@narfu.ru

About the authors

Berendyaev Maxim Viktorovich – Associate Professor, Department of Translation Technology and Practice at AKM-WEST, Northern (Arctic) Federal University, Russia, Arkhangelsk, m.berendyaev@narfu.ru

Gilin Mikhail Igorevich – Assistant Professor, Department of Foreign Languages and Communication Technologies, College of Basic Professional Studies, MISIS University of Science and Technology, Russia, Moscow, m.gilin@misis.ru

Kokanova Elena Sergeevna – Ph. D. of Philology, Docent, Head of Department of Translation Technology and Practice at AKM-WEST, Northern (Arctic) Federal University, Russia, Arkhangelsk, e.s.kokanova@narfu.ru