ISSN 2658-5014 (Print) СИИТ ISSN 2686-7044 (Online)
2025. Т. 7, № 2 (21). С. 30–47 http://siit.ugatu.su
СИСТЕМНАЯ ИНЖЕНЕРИЯ И ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ
Научная статья

УДК 004.89 DOI 10.54708/2658-5014-SIIT-2025-no2-p30 EDN BWYTSX

РАЗРАБОТКА МОДЕЛИ НЕЙРОННОГО МАШИННОГО ПЕРЕВОДА ДЛЯ МАНСИЙСКОГО ЯЗЫКА

О. О. НЕГМАТУЛОЕВ • Д. О. ЖОРНИК • А. В. МЕЛЬНИКОВ

Аннотация. В работе представлено описание процесса обучения трансформерной нейронной сети для решения задачи машинного перевода для мансийского языка (< обско-угорские < финно-угорские < уральские), являющегося в настоящее время малоресурсным. Целью работы является проведение экспериментов по сравнению результатов дообучения мультиязычных моделей для языковой пары: русского и мансийского языков. В работе приведен обзор современных методов машинного перевода и архитектур нейронных сетей, включая трансформерные сети. В результате работы были дообучены нейронные сети с использованием библиотек PyTorch и Transformers. Качество перевода оценивалось метриками BLEU и chrF. Лучший результат был получен для модели NLLB-200-3.3B, которая достигла показателей BLEU 27 % и chrF 57 % для перевода с русского на мансийский язык. Проведены дополнительные эксперименты и анализ для выявления сильных и слабых сторон методов с помощью экспертной оценки. Работа демонстрирует эффективность применения трансформерных моделей в задаче машинного перевода и может быть использована в практических приложениях.

Ключевые слова: малоресурсные языки; машинный перевод; финно-угорские языки; мансийский язык.

ВВЕДЕНИЕ

Машинный перевод является одним из ключевых направлений в искусственном интеллекте и обработке естественного языка. За последние годы благодаря быстрому развитию методов машинного обучения, особенно нейронных сетей, качество автоматического перевода значительно возросло. Появление крупных языковых моделей, способных генерировать текст на уровне человека, открыло новые возможности для преодоления языковых барьеров и укрепления межкультурного взаимодействия.

Создание моделей нейронного машинного перевода основывается на использовании больших параллельных корпусов, что позволяет достичь высокого качества перевода для языковых пар с достаточным количеством обучающих данных. Обучаясь на таких обширных параллельных ресурсах, модели способны учитывать сложные языковые структуры, семантику и контекст, что делает их переводы практически сопоставимыми с работой человека. Наличие богатых данных позволяет моделям эффективно обрабатывать широкий спектр языковых особенностей и вариаций, присущих как исходным, так и целевым языкам.

Однако развитие машинного перевода для малоресурсных языков сопровождается рядом сложностей. Эти языки, как правило, имеют ограниченные объемы доступных параллельных текстов, что затрудняет использование традиционных подходов нейронного машинного перевода, требующих большого количества данных. Для решения этой проблемы исследователи разрабатывают современные методы, такие как трансферное обучение и многоязычные модели. Эти подходы позволяют адаптировать нейронные модели к малоресурсным языкам, используя знания, полученные при обучении на языках с богатой базой данных.

Мансийский язык относится к обско-угорской подгруппе финно-угорской ветви уральской языковой семьи и распространен на территории Российской Федерации, в первую очередь, в Ханты-Мансийском автономном округе, а также в Свердловской области. Количество носителей языка по последним данным (см., например, Перепись 2020/2021) оценивается в районе 1 000 человек, и в подавляющем большинстве это носители среднего и старшего поколения. Таким образом, мансийский язык находится под серьезной угрозой исчезновения,

и в таких условиях создание компьютерных инструментов представляется особенно актуальным, поскольку расширит возможности изучения языка заинтересованными людьми и повысит его представленность в общественном пространстве. Кроме того, со стороны самого мансийского сообщества выражается активный интерес к разработке новых ресурсов и материалов для поддержания и развития мансийского языка.

Таким образом, даже для языков с ограниченными ресурсами становится возможным создавать эффективные системы машинного перевода. Целью данной работы является разработка системы нейронного машинного перевода для мансийского языка.

Описание методов

Тема нейронного машинного перевода для малоресурсных языков привлекает значительное внимание исследователей в области обработки естественного языка. Ряд работ предлагают различные подходы к решению этой задачи.

Трансферное обучение для машинного перевода [1, 2]. Некоторые исследования фокусируются на использовании трансферного обучения, где модели, предварительно обученные на богатых ресурсами языках, адаптируются к малоресурсным языкам. Например, авторы статьи «Transfer Learning for Low-Resource Neural Machine Translation» [1] показали, что такой подход может значительно улучшить качество перевода для языков с ограниченными ресурсами. Используя метод трансферного обучения, они смогли улучшить базовые модели нейронного машинного перевода в среднем на 5.6 % BLEU (Bilingual Evaluation Understudy) для четырех языковых пар с ограниченными ресурсами.

Многоязычные модели [3–7]. Другие работы исследуют использование многоязычных нейронных сетевых моделей, которые могут совместно обучаться на нескольких языках. Такие модели, как правило, демонстрируют лучшую производительность на малоресурсных языках по сравнению с моделями, обученными только на одном языке. На данный момент вышло много предобученных мультиязыковых моделей, которые владеют более 200 языками. Например, модель NLLB [8] (No Language Left Behind «Ни один язык не будет забыт») была разработана для расширения возможностей машинного перевода на широкий спектр языков, особенно тех, которые традиционно считались «языками с ограниченными ресурсами». Модель NLLB охватывает более 200 языков, что значительно превосходит языковой охват традиционных систем машинного перевода. Другим примером является семейство моделей MADLAD400 [9] от компании Google, которые способны переводить на более 400 языков. Такие предварительно обученные языковые модели требуют гораздо меньше времени на обучение по сравнению с моделями, обучаемыми с нуля, и их можно использовать для дообучения под конкретные задачи.

Ключевым преимуществом использования предварительно обученных моделей является их способность достигать высокого качества перевода даже для языков с очень ограниченными параллельными данными. Также многоязычные модели могут содержать в себе родственные языки для целевого языка, что также может улучшить качество машинного перевода. Включение родственных языков в обучение многоязычной модели позволяет использовать лингвистические связи и сходства между этими языками для более эффективной передачи знаний и улучшения обобщения. Это открывает новые возможности для расширения доступности машинного перевода на широкий спектр малоресурсных языков.

Метрики оценки качества машинного перевода

Оценка качества машинного перевода является важной и сложной задачей. Одним из примеров является ежегодная конференция WMT (Workshop on Machine Translation), в рамках которой одной из ключевых задач является поиск лучших метрик, коррелирующих с оценками, выставленными экспертами. Ежегодно на конференции WMT проводится «Shared Task» по оценке метрик машинного перевода. Участники предоставляют автоматические метрики

оценки качества перевода, которые затем оцениваются по степени их корреляции с человеческими оценками на уровне систем и отдельных предложений. Для получения человеческих оценок используется экспертная оценка с применением многомерных метрик качества (Multidimensional Quality Metrics, MQM). Существуют ряд метрик, которые широко используются для этой цели.

Автоматические метрики

Лексические метрики для оценки качества машинного перевода позволяют количественно измерить, насколько хорошо переведенный текст соответствует эталонному переводу на уровне слов и фраз. Вот некоторые из наиболее распространенных лексических метрик:

BLEU (Bilingual Evaluation Understudy) [10] – это одна из наиболее распространенных метрик для оценки машинного перевода. BLEU измеряет степень совпадения *n*-грамм в переводе с эталонными переводами, выполненными человеком. Более высокие значения BLEU указывают на лучшее качество перевода.

METEOR (Metric for Evaluation of Translation with Explicit Ordering) [11] – в отличие от BLEU, METEOR учитывает не только точность *n*-грамм, но и их семантическое соответствие. Она также принимает во внимание синонимы и парафразы, что делает ее более чувствительной к качеству перевода. В задаче машинного перевода METEOR использует такие ресурсы, как синонимические словари и парафразы, для более глубокой оценки. Для малоресурсных языков доступность и качество таких ресурсов ограничены, что снижает надежность METEOR.

TER (Translation Edit Rate) [12] – эта метрика оценивает количество редактирований (вставок, удалений, замен и перестановок слов), необходимых для преобразования машинного перевода в эталонный. Более низкие значения TER соответствуют лучшему качеству перевода.

chrF (Character n-gram F-score) [13] – данная метрика основана на совпадении символьных *n*-грамм между переводом и эталоном. Она показала хорошую корреляцию с человеческой оценкой качества перевода, особенно – для языков с богатой морфологией.

Традиционные метрики оценки качества текста, такие как BLEU, часто критикуются за их ограниченную способность улавливать семантическое сходство между текстами. Эти метрики в основном основаны на сопоставлении *n*-грамм, что не всегда отражает глубокое понимание языка. Однако в последние годы появился новый класс метрик, которые преодолевают эти ограничения, используя векторные представления слов и фраз. Эти метрики способны уловить более тонкие семантические связи между текстами, что делает их более подходящими для оценки качества текстовой генерации и понимания. В этой главе мы рассмотрим несколько ключевых метрик:

BertScore [14] – данная метрика основывается на векторных представлениях слов и фраз, на основе предобученных моделей, таких как BERT (Bidirectional Encoder Representations from Transformers), Roberta.

COMET (Crosslingual Optimized Metric for Evaluation of Translation) [15]. Данная метрика так же, как и BERTScore, основывается на векторных представлениях слов и фраз, на основе предобученных моделей. Но в отличие от BertScore, COMET включает информацию как из исходного текста, так и из эталонного перевода на целевой язык, чтобы более точно прогнозировать качество машинного перевода.

Экспертные метрики

Экспертная оценка машинного перевода. Для экспертной оценки машинного перевода на конференции WMT используется многомерная метрика качества. MQM (Multidimensional Quality Metrics) [16–20] – это система оценки качества перевода, которая позволяет пользователям настраивать собственные метрики для оценки качества. Она была разработана для создания общей системы оценки качества как для человеческого, так и для машинного перевода, а также для улучшения автоматической оценки качества перевода и на данный момент различает более 100 видов ошибок машинного перевода.

MQM была создана, чтобы предоставить унифицированную систему оценки качества, которую можно адаптировать под конкретные потребности пользователей. Она позволяет настраивать метрики оценки в зависимости от типа контента, целевой аудитории и других факторов. Основная цель MQM – улучшить оценку качества как человеческого, так и машинного перевода и помочь повысить общее качество автоматического перевода.

Центральным компонентом MQM является иерархический список типов проблем. Типы были получены в результате тщательного изучения существующих метрик оценки качества и проблем, выявляемых автоматическими инструментами проверки качества.

В рамках обучения модели машинного перевода важно учитывать различные типы ошибок, которые могут возникать в переводах. Ниже приведено краткое описание семи основных типов ошибок (табл. 1).

Таблица 1

Иерархия основных ошибок MQM

Категория	Ошибка	Описание
Accuracy (Точность)	Addition (Добавление)	Перевод включает в себя информацию, которой не было в источнике
	Omission (Упущение)	В переводе пропущена информация, находящаяся во входном тексте
	Mistranslation (Искажение)	Перевод не точно передает информацию, которая была во входном тексте
	Untranslated (Не переведен)	Входной текст не был переведен
Fluency (Связность)	Punctuation (Пунктуация)	Неверная пунктуация
	Spelling (Орфография)	Неверное написание или использование заглавных букв
	Grammar (Грамматика)	Проблемы с грамматикой
Terminology (Терминология)	Inappropriate for context	Терминология нестандартна или не соответствует содержанию
	Inconsistent use	Терминология использована не верно
Style	Awkward	Проблемы со стилистикой перевода
Source Error		Ошибка в исходном предложении
Non translation		Невозможно достоверно охарактеризовать различимую ошибку
Other		Другой вид ошибки

Также выделяются уровни серьезности ошибок перевода: незначительные (Minor: не вводят в заблуждение и не меняют значение перевода), значительные (Major: меняют значение перевода). Выделяется и третий тип – критические (Critical: меняют значение перевода и несут какие-либо последствия, возможно, оскорбительные), но данный тип обычно не используется.

Помимо этой иерархической системы ошибок, на конференциях WMT также применялась скалярная метрика качества (Scalar Quality Metric, SQM) для оценки машинного перевода. В этом подходе предлагается оценивать качество перевода по шкале от 0 до 6 баллов.

Выбор основных метрик для мансийского языка

Исходя из результатов конференции WMT 2023 Shared Tasks [21] (рис. 1) по оценке метрик машинного перевода можно сделать несколько важных выводов. Одним из ключевых наблюдений является то, что метрики, основанные на векторных представлениях текста и слов, демонстрируют наиболее высокую корреляцию с оценками экспертов.

Metric		avg corr
XCOMET-Ensemble	**1**	**0.825**
XCOMET-QE-Ensemble*	2	0.808
MetricX-23	2	0.808
GEMBA-MQM*	2	0.802
MetricX-23-QE*	2	0.800
mbr-metricx-qe*	3	0.788
MaTESe	3	0.782
CometKiwi*	3	0.782
COMET	3	0.779
BLEURT-20	3	0.776
KG-BERTScore*	3	0.774
sescoreX	3	0.772
cometoid22-wmt22*	4	0.772
docWMT22CometDA	4	0.768
docWMT22CometKiwiDA*	4	0.767
Calibri-COMET22	4	0.767
Calibri-COMET22-QE*	4	0.755
YiSi-1	4	0.754
MS-COMET-QE-22*	5	0.744
prismRef	5	0.744
mre-score-labse-regular	5	0.743
BERTscore	5	0.742
XLsim	6	0.719
f200spBLEU	7	0.704
MEE4	7	0.704
tokengram_F	7	0.703
embed_llama	7	0.701
BLEU	7	0.696
chrF	7	0.694
eBLEU	7	0.692
Random-sysname*	8	0.529
prismSrc*	9	0.455

Рис. 1 Результаты WMT2023 Shared Tasks

Традиционные метрики, такие как BLEU, которые основаны на сравнении *n*-грамм слов, показали более низкую корреляцию с человеческими оценками. В то же время метрики, использующие более сложные векторные представления (COMET, BertScore), такие как семантическое сходство и контекстуальная близость, оказались более эффективными в отражении качества перевода, воспринимаемого человеком.

В целом тенденции, выявленные на конференции WMT 2023, подчеркивают важность использования более продвинутых подходов к оценке машинного перевода, основанных на векторных представлениях, для получения достоверных и содержательных результатов.

В данной работе не предполагается использование метрик, основанных на векторных представлениях текста или слов, так как такие подходы требуют наличия предварительно обученных языковых моделей, способных создавать качественные векторные представления. В настоящее время для мансийского языка, который относится к малоресурсным и является одним из объектов исследования, подобной языковой модели не существует. Поэтому в рамках исследования будут применяться более традиционные метрики для оценки качества перевода.

Кроме того, значительное внимание будет уделено экспертной оценке качества перевода носителями языка. Несмотря на высокую трудоемкость, такой подход позволяет учитывать различные аспекты перевода, включая его адекватность, естественность и понятность.

На первоначальном этапе обучения моделей, с учетом особенностей мансийского языка, будут использоваться автоматические метрики, такие как BLEU и chrF. После завершения обучения на основе этих метрик будут отобраны лучшие модели, которые затем пройдут дополнительную оценку качества с привлечением экспертов по адаптированной методике MQM.

Адаптация MQM для мансийского языка

Методика MQM была адаптирована исходя из доступных ресурсов и поставленной задачи. Были оставлены два уровня серьезности ошибок «Серьезная» и «Не серьезная» (табл. 2). Также мы выбрали три категории ошибок – это точность, связность, стиль, и оставили ошибки «Ошибка в эталоне» и «Ошибка в источнике» (табл. 3).

Таблица 2

Описание уровней серьезности ошибок

Серьезность ошибки	Описание
Серьезная	Меняет смысл перевода и вводит в заблуждение
Не серьезная	Не меняет смысл перевода

Таблица 3

Описание иерархии ошибок

Категория ошибки	Описание категории	Название ошибки	Описание ошибки
Точность	Оценивается правильность передачи смысла исходного текста	Добавление	В переводе присутствует новая информация, которой не было в исходном тексте
		Упущение	В переводе отсутствует информация, которая присутствовала в исходном тексте
		Искажение	Перевод неточно передает смысл исходного текста, используются неправильные эквиваленты
		Непереведенный текст	Фрагмент исходного текста остался непереведенным и присутствует в оригинальном виде
		Лексика	Неверно подобрано слово или оно используется не в том значении
Связность	Оценивается связность и понятность текста перевода с точки зрения языковых норм	Пунктуационные	Ошибки в расстановке знаков препинания, затрудняющие понимание текста
		Орфографические	Ошибки в написании слов, не соответствующие правилам орфографии языка перевода
		Грамматические	Ошибки в употреблении грамматических форм, конструкций и согласовании, нарушающие языковые нормы
Стиль	Оценивается соответствие перевода стилистическим нормам и особенностям языка перевода	Стилистические	Перевод не соответствует стилистическим нормам языка перевода, содержит элементы, неуместные для данного типа текста
		Неестественность	Перевод звучит неестественно для носителя языка, хотя и не содержит явных ошибок
Дополнительные ошибки		Ошибка в источнике	Ошибка в исходном предложении
		Ошибка в эталоне	Ошибка в переводе, сделанном экспертом

НАБОР ДАННЫХ

Одним из ключевых факторов, влияющих на качество нейронного машинного перевода, является наличие и качество параллельных корпусов текстов, используемых для обучения. Для мансийского языка на данный момент не существует достаточного корпуса параллельных предложений в современной графике мансийского языка.

Сбор корпуса параллельных предложений

Сбор данных был организован с использованием специально разработанного сервиса для создания параллельных переводов. Для этого сначала были агрегированы материалы на мансийском языке (тексты мансийской газеты «Лӯимā Сэ̄рипос», переводы текстов Библии на мансийском языке, материалы из фольклорных сборников), которые были переведены в машиночитаемый вид и загружены в программу для дальнейшей обработки, то есть формирования перевода.

Переводчик получает предложение из предложенного набора или может ввести новое предложение вручную. Если переводчик получил предложение, которое он не может перевести, он может его пропустить, и тогда это предложение попадает в конец очереди. Когда переводчик перевел предложение, он отправляет его на проверку эксперту. Сервис выдает случайному эксперту предложение и его перевод, при этом эксперт не знает автора перевода, что повышает качество оценки за счет объективности и непредвзятости.

Эксперт может принять перевод, при этом, если есть небольшие ошибки, он может сам поправить перевод и утвердить его. Если же предложение переведено «плохо», то эксперт вернет предложение на доработку переводчику, при этом может оставить свой комментарий, уточнив, какие ошибки допустил переводчик. На данный момент собранный корпус является первым масштабным проектом для мансийского языка.

Информация по собранному набору данных

На момент проведения экспериментов по обучению модели машинного перевода был создан параллельный корпус, включающий около 55 тысяч предложений для обучения модели машинного перевода с мансийского на русский язык. Со временем объем набора увеличился сначала ~ до 120 тысяч, и для него будет описан отдельный эксперимент по обучению модели. Однако в данных было выявлено несколько проблем, которые, хотя и не являются критическими, требуют внимания:

- неправильное разделение слов пробелами: в некоторых случаях слова были разделены некорректно (например, "с л о в о"), что усложняет обработку текста;
- различия в обозначении долготы звуков: в мансийском языке различаются долгие и краткие гласные, и в некоторых ситуациях это различие имеет смыслоразличительный характер (например, тур ‘горло’ vs. тӯр ‘озеро’), однако среди носителей мансийского языка есть широкая вариативность относительно использования долгот на письме. В корпусе также наблюдалась непоследовательность в использовании символов для обозначения долготы звуков, что могло приводить к ошибкам в распознавании и переводе;
- ошибки в направлении перевода: были случаи, когда вместо мансийского текста в паре использовался русский, и наоборот, что могло сбивать модель с толку;
- кроме того, в процессе анализа результатов обучения и сравнения машинного и человеческого переводов была выявлена еще одна интересная особенность. Предложения, переведенные профессиональными переводчиками, зачастую были вырваны из контекста, и их перевод основывался на предполагаемых значениях. Модель же, не имея доступа к контексту, выполняла перевод буквально. Эксперты отметили, что такой подход также можно считать корректным в части случаев.

Для решения указанных проблем были разработаны специальные скрипты для предварительной обработки данных. Эти скрипты позволили частично или полностью устранить большинство недостатков и значительно улучшить качество перевода.

В целом благодаря продуманному процессу сбора параллельного корпуса удалось создать достаточно качественный набор данных для обучения модели. Это позволило подготовить надежную основу для перехода к следующему этапу обучения.

Обучение модели

После подготовки корпуса параллельных предложений для мансийского и русского языков следующим шагом является обучение нейронной сети для задачи машинного перевода. В данной работе был использован подход нейронного машинного перевода на основе архитектуры трансформерной нейронной сети.

Технологии, использованные в обучении

Для разработки и обучения нейронной системы машинного перевода с мансийского на русский язык были использованы следующие технологии и инструменты:

- PyTorch: Популярная библиотека для работы с машинным обучением, обеспечивающая широкий набор инструментов для создания и обучения нейронных сетей. PyTorch был выбран благодаря своей гибкости, удобству использования и активному сообществу разработчиков. В работе использовалась версия с поддержкой CUDA 12;
- Transformers: библиотека от компании HuggingFace, предоставляющая готовые предварительно обученные модели трансформеров для задач обработки естественного языка, включая задачи машинного перевода;
- ClearML: платформа для управления процессом машинного обучения, применявшаяся для отслеживания параметров обучения, метрик, версий моделей и других артефактов. ClearML позволила оптимизировать процесс разработки и обеспечить воспроизводимость экспериментов;
- система контроля версий: исходный код проекта, включая скрипты для обучения, предобработки данных и вспомогательные утилиты, хранился и управлялся с использованием GitLab;
- язык программирования: проект был реализован на Python версии 3.11. Для управления зависимостями и настройкой среды использовался менеджер пакетов Poetry.

Для обучения и тестирования нейронной сети была задействована следующая вычислительная инфраструктура: модель обучалась на выделенном вычислительном кластере, оснащённом 6 GPU NVIDIA A100 с объёмом видеопамяти 80 Гб каждая.

Предобученные модели

В качестве предобученных моделей использовались следующие:

- MADLAD-400-3b;
- MADLAD-400-7b;
- MADLAD-400-10b;
- NLLB-200-3.3b.

MADLAD400 и NLLB200 использует SentencePiece токенизатор. SentencePiece работает следующим образом [22]. Входной текст рассматривается как непрерывная последовательность символов Unicode, включая пробелы и знаки препинания. Пробелы заменяются специальным символом "▁", который используется для обозначения границ слов. Затем происходит построение словаря. SentencePiece использует алгоритм BPE (Byte Pair Encoding) или Unigram для построения словаря подслов (subword). Алгоритм итеративно объединяет часто встречающиеся пары символов (для BPE) или отдельные символы (для Unigram) в новые подслова. Процесс продолжается до достижения заданного размера словаря. Процесс токенизации происходит следующим образом: входной текст разбивается на последовательность подслов из построенного словаря. Каждое подслово заменяется соответствующим идентификатором из словаря. Если слово не может быть представлено подсловами из словаря, оно разбивается на отдельные символы.

Таблица 4

Информация о токенизаторах

Название модели	Тип токенизатора	Размер словаря (токен)
google/madlad400	SentencePiece	256 000
facebook/nllb-200	SentencePiece	256 204 (включая языковые коды)

Чтобы оценить качество токенизации текста на мансийском языке, были проведены следующие эксперименты [23]:

- эксперимент проводился на ~55 тысячах параллельных предложениях;
- предложения из корпуса мансийского языка токенизированы с помощью предобученного токенизатора;
- подсчитано количество предложений, содержащих специальный токен «UNK», который обозначает неизвестные токены (табл. 5);
- также было подсчитано среднее количество токенов на слово *русский* для русского и мансийского языков.

Таблица 5

Результаты эксперимента по качеству токенизации

Название токенизатора	Количество предложений, в которых встретился токен UNK (предложения)
MADLAD400	32
NLLB200	22 424

Таблица 6

Результаты эксперимента по качеству токенизации

Язык	NLLB200 (среднее количество токенов на слово)	MADLAD400 (среднее количество токенов на слово)
Русский	~1.68	~1.69
Мансийский	~2.66	~2.53

На основании данных из табл. 5 и 6 можно сделать вывод, что в среднем на одно мансийское слово приходится более двух токенов. Это может свидетельствовать о сложностях, связанных с обработкой мансийского языка.

Результаты показывают, что модель NLLB-200 испытывает трудности с токенизацией текстов на мансийском языке. Количество предложений, содержащих неизвестные токены, составляет 22 424, что значительно больше, чем у модели MADLAD400, у которой таких предложений всего 32. Основной причиной этого является недостаточный размер словаря модели NLLB-200, который не охватывает все особенности мансийского языка, в частности, в словаре отсутствует символ «ӈ».

Для модели NLLB-200 токенизация текстов показала, что значительное количество предложений содержит токен UNK. Это указывает на ограничения модели в обработке мансийского языка. Примеры токенизации, иллюстрирующие эту проблему, представлены в табл. 7.

Для устранения этой проблемы было предложено расширить словарь модели NLLB-200, добавив в него всю необходимую символику и токены мансийского языка.

Таблица 7

Примеры токенизации на базовом токенизаторе предобученной модели

	Русский	Мансийский
Оригинал	Сейчас я хочу показать вам три варианта развития	Ам ань таӈхēгум нāнан хӯрум йильпииг вāрнэ тэ̄ла сунсылтаӈкве.
MADLAD400	['▁Сей', 'час', '▁я', '▁', 'хочу', '▁по', 'казать', '▁вам', '▁три', '▁вариант', 'а', '▁развития', '.', '</s>']	['▁Ам', '▁', 'ань', '▁та', 'ӈ', 'х', 'ē', 'гум', '▁на', '̄', 'нан', '▁', 'хӯр', 'ум', '▁йил', 'ь', 'пи', 'иг', '▁ва', '̄', 'р', 'нэ', '▁тэ', '̄', 'ла', '▁', 'сун', 'сыл', 'та', 'ӈ', 'кве', '.', '</s>']
NLLB200	['rus_Cyrl', '▁Сейчас', '▁я', '▁хочу', '▁показа', 'ть', '▁вам', '▁три', '▁вари', 'анта', '▁развития', '.', '</s>']	['mansi_Cyrl','▁Ам', '▁ан', 'ь', '▁та', '<unk>', 'х', 'ē', 'гу', 'м', '▁на', '̄', 'нан', '▁хӯ', 'рум', '▁й', 'и', 'ль', 'пи', 'и', 'г', '▁ва', '̄', 'р', 'нэ', '▁тэ', '̄', 'ла', '▁сун', 'сыл', 'та', '<unk>', 'кве', '.', '</s>']

Для модели MADLAD400 будет также проведено расширение словаря для разбиения текста на более осмысленные слова, а не части слов и отдельные символы. Процесс расширения токенизатора включал следующие шаги:

- анализ токенизатора: был проведен анализ работы токенизатора для мансийской символики и выявлены необходимые символы для полного покрытия токенизатора мансийскими символами;
- создание расширенного словаря: в словарь базового токенизатора были добавлены необходимые символы и слова мансийского языка;
- был проведён процесс дообучения токенизатора на корпусе текстов мансийского языка;
- после обучения токенизатора была проведена ручная чистка словаря обученной модели. В процессе чистки словарь был очищен от лишних токенов, содержащих различные знаки препинания; избыточные токены, включающие именованные сущности, такие как имена собственные, названия географических объектов и другие элементы.

После расширения словаря моделей среднее количество токенов на слово было ~1.5. Пример токенизации на расширенном токенизаторе предобученной модели представлен в табл. 8.

Таблица 8

Пример результата токенизации на расширенном токенизаторе

	Русский	Мансийский
Оригинал	Сейчас я хочу показать вам три варианта развития	Ам ань таӈхēгум нāнан хӯрум йильпииг вāрнэ тэ̄ла сунсылтаӈкве.
MADLAD400	['▁Сейчас', '▁я', '▁', 'хочу', '▁п', 'ока', '▁за', 'ть', '▁вам', '▁три', '▁вариант', 'а', '▁', 'развития', '.', '▁', '</s>']	['▁Ам', '▁', 'ань', '▁', 'таӈ', '▁х', 'ēгум', '▁', 'нāн', '▁', 'ан', '▁', 'хӯрум', '▁', 'йильпи', '▁', 'иг', '▁', 'вāр', '▁нэ', '▁', 'тэ̄ла', '▁су', 'нс', 'ылтаӈкве', '▁', '.', '▁', '</s>']
NLLB200	['rus_Cyrl', '▁Сейчас', '▁я', '▁хочу', '▁показа', 'ть', '▁вам', '▁три', '▁вари', 'анта', '▁развития', '.', '</s>']	['mansi_Cyrl','▁Ам', '▁ань', '▁таӈхēгум', '▁нāнан', '▁хӯрум', '▁йильпииг', '▁вāрнэ', '▁тэ̄ла', '▁сунсыл', 'таӈкве', '.' '</s>']

Обучение трансформерной нейронной сети

Для обучения был использован датасет параллельных состоящий из ~55 тысяч параллельных предложений, он был разделен на оценочную и обучающую выборки, в следующем соотношении:

- обучающая 41 682 пары предложений;
- оценочная 13 894 пары предложений.

Для обучения модели были заданы следующие параметры:

- оптимизатор – для обновления параметров модели использовался алгоритм оптимизации AdamW;
- скорость обучения (Learning Rate) – начальное значение было установлено на уровне 2e-5. Этот параметр изменялся динамически в процессе обучения;
- количество эпох – модель обучалась в течение 6 эпох, каждая из которых представляет собой полный проход по всему обучающему датасету;
- контрольные точки (Checkpoints): для экономии дискового пространства на сервере сохранялась только одна контрольная точка. Контрольная точка представляет собой сохраненные веса модели на определенном этапе обучения и фиксировалась после каждого шага (step), кратного заданному интервалу. Один step соответствует обработке моделью одного пакета данных;
- функция потерь (Loss) – CrossEntropy – в контексте языковых моделей это означает сравнение вероятностей, присвоенных моделью каждому токену в последовательности, с фактическими токенами в обучающем наборе данных.

Для оценки качества перевода были использованы следующие параметры метрик:

- BLEU: вычислялся с учетом до 4 грамм (4-ngrams);
- chrF: рассчитывался с учетом до 6 грамм (6-ngrams).

Чтобы многоязычные модели понимали, с какого языка на какой выполнять перевод, были использованы специальные «префиксы». Для семейства моделей madlad400 были назначены следующие префиксы:

- <ru2mansi> – для перевода с русского на мансийский;
- <mansi2ru> – для перевода с мансийского на русский.

Для модели nllb-200 был добавлен специальный языковой код (mansi_Cyrl) для обозначения мансийского языка, данные языковые коды будут являться BOS (Begin of the text) токеном и будут выставлены в начале предложения.

Эти обозначения позволяют явно указывать направление перевода и обеспечивают корректную работу многоязычной модели. Таким образом, модель может понимать, какое преобразование требуется выполнить – перевод с русского на мансийский или наоборот. Использование таких обозначений является важной деталью, позволяющей многоязычным моделям эффективно работать с различными парами языков для перевода.

Результаты обучения на тестовой выборке представлены в табл. 9.

Таблица 9

Результаты обучения моделей

	BLEU	ChrF
	ru→mansi	ru→mansi
google/madlad400-3b-mt	12.1	36.59
google/madlad400-7b-mt-bt	12.6	37.23
google/madlad400-10b-mt	14.2	39.2
facebook/nllb200-3.3B	**23.2**	**52.06**

На рис. 2 показаны график роста BLEU метрики и уменьшение ошибки (loss) на тестовом датасете в течение 6 эпох для модели MADLAD400, данное поведение графиков также свойственно модели NLLB.

Следующим экспериментом было обучение модели NLLB, так как она лучше показала себя по автоматическим метрикам на наборе данных, содержащим ~120 тысяч параллельных предложений.

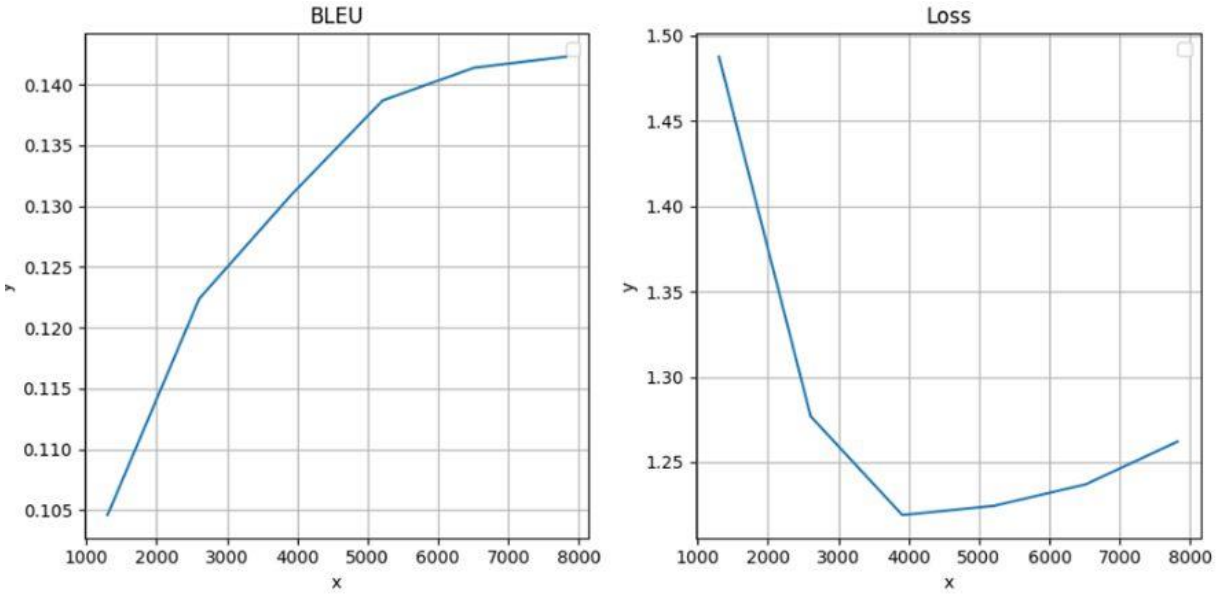

Рис. 2 Графики оценок

Результаты обучения на расширенном наборе данных

Здесь будут показаны результаты обучения модели на расширенном корпусе параллельных данных, который содержал в себе 120 тысяч параллельных предложений.

Учитывая ограниченность данных, было принято решение скорректировать соотношение выборок. Вместо выделения 20 % набора данных на оценочную выборку, что является избыточным в условиях малых данных, большая часть была направлена на обучающую выборку. Помимо этого, количество эпох для обучения модели было увеличено с 6 до 12. В итоге было обучено две модели для достижения двунаправленного перевода.

Итоговый размер обучающей и тестовой выборки:

- тестовая выборка – 2418 пар предложений;
- обучающая выборка – 121 365 пар предложений.

Все параметры эксперимента остались неизменными по сравнению с предыдущими экспериментами, за исключением используемого числа эпох обучения. В рамках данного эксперимента модель обучалась в течение 12 эпох для модели с русского на мансийский, так как ошибка на тестовой выборке продолжала падать, и после 6 эпохи метрика BLEU уже выходила на плато, для модели с русского на мансийский модель была обучена на 5 по таким же соображениям. Результаты лучшего обучения представлены в табл. 10 и 11.

Таблица 10

Результаты оценок качества перевода модели с русского на мансийский

	BLEU	ChrF
	ru→mansi	ru→mansi
facebook/nllb200-3.3B	27.3	56.7

Таблица 11

Результаты оценок качества перевода модели с мансийского на русский

	BLEU	ChrF
	mansi→ru	mansi→ru
facebook/nllb200-3.3B	25.3	50.3

Это был последний эксперимент по обучению модели машинного перевода для языковых пар русский мансийский с использованием автоматических метрик оценки качества. На автоматических метриках удалось добиться 27.3 % для переводов с русского на мансийский и 25.3 % для переводов с мансийского на русский, что для данных метрик является неплохим результатом. Также важно помнить о том, что основная функция этих метрик – сравнение производительности различных моделей и отслеживание динамики улучшений. Они дают возможность оценить относительные улучшения модели по сравнению с предыдущими версиями или другими подходами. Однако окончательная оценка качества модели должна основываться на экспертном анализе и человеческой оценке, которые позволяют более глубоко проанализировать смысловую точность и качество перевода.

Экспертная оценка обученных моделей

После обучения модели на 120 тысячах пар параллельных предложений была проведена экспертная оценка с использованием методики, описанной в разделе «Адаптация MQM для мансийского языка». Для проведения оценки был отобран датасет размером 2 тысячи пар параллельных предложений перевода с русского на мансийский язык и 2 тысячи пар параллельных предложений перевода с мансийского на русский язык. В процессе оценки эксперт имел возможность определять уровень серьезности каждой ошибки, обнаруженной в предложениях.

На рис. 3 представлено распределение предложений с ошибками и без ошибок. Согласно данным, около 77 % предложений оказались без ошибок для переводов с русского на мансийский. На рис. 4 представлено соотношение ошибок по уровню серьезности среди предложений с ошибками. Из него видно, что среди предложений с ошибками всего лишь 10.4 % ошибок являются серьезными.

Рис. 3 Соотношение предложений с ошибками и без ошибок (с русского на мансийский)

Рис. 4 Соотношение серьезных и несерьезных ошибок (с русского на мансийский)

На рис. 5 представлено распределение ошибок по категориям, что позволяет выявить области с наибольшим количеством ошибок. При переводах с русского языка часто встречаются грамматические и лексические ошибки, а также искажения и неестественные формулировки. На рис. 6 изображено распределение ошибок по категориям, а также уровням серьезности ошибок. На этом графике можно увидеть, в каких категориях чаще допускаются ошибки по уровню серьезности. Из этого распределения можно сделать вывод, что наиболее критические ошибки совершаются в категориях Искажение и Лексика. На рис. 7 изображено соотношение предложений с ошибками и без для модели с мансийского на русский. Здесь уже количество предложений с ошибками больше 50 %, множество ошибок были связаны с орфографией, модель совершала множество орфографических ошибок на русском языке, что привело к значительному ухудшению перевода. На рис. 7–10 представлены аналогичные сведения для перевода с мансийского на русский.

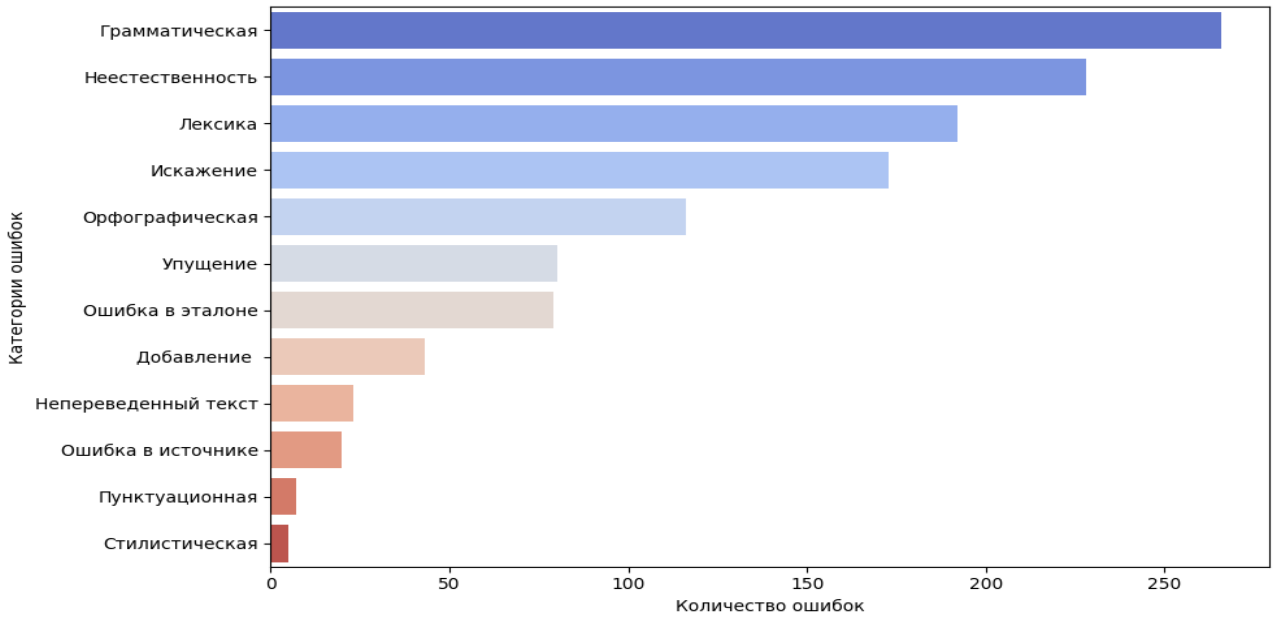

Рис. 5 Распределение ошибок по категориям (с русского на мансийский)

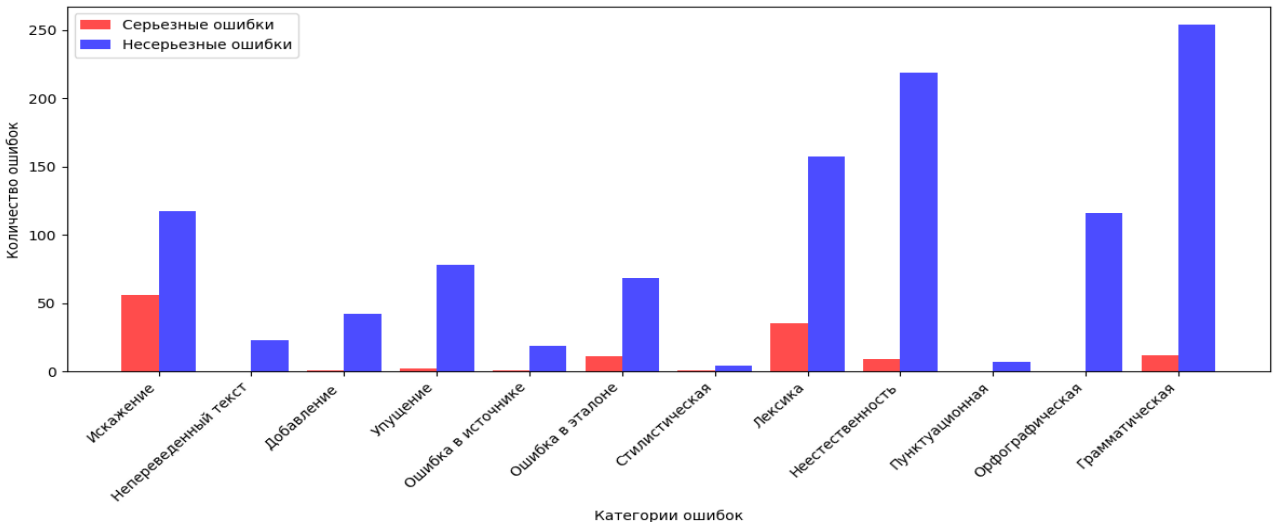

Рис. 6 Распределение ошибок по категориям и уровням серьезности (с русского на мансийский)

Рис. 7 Соотношение предложений с ошибками и без ошибок (с мансийского на русский)

Рис. 8 Соотношение серьезных и несерьезных ошибок (с мансийского на русский)

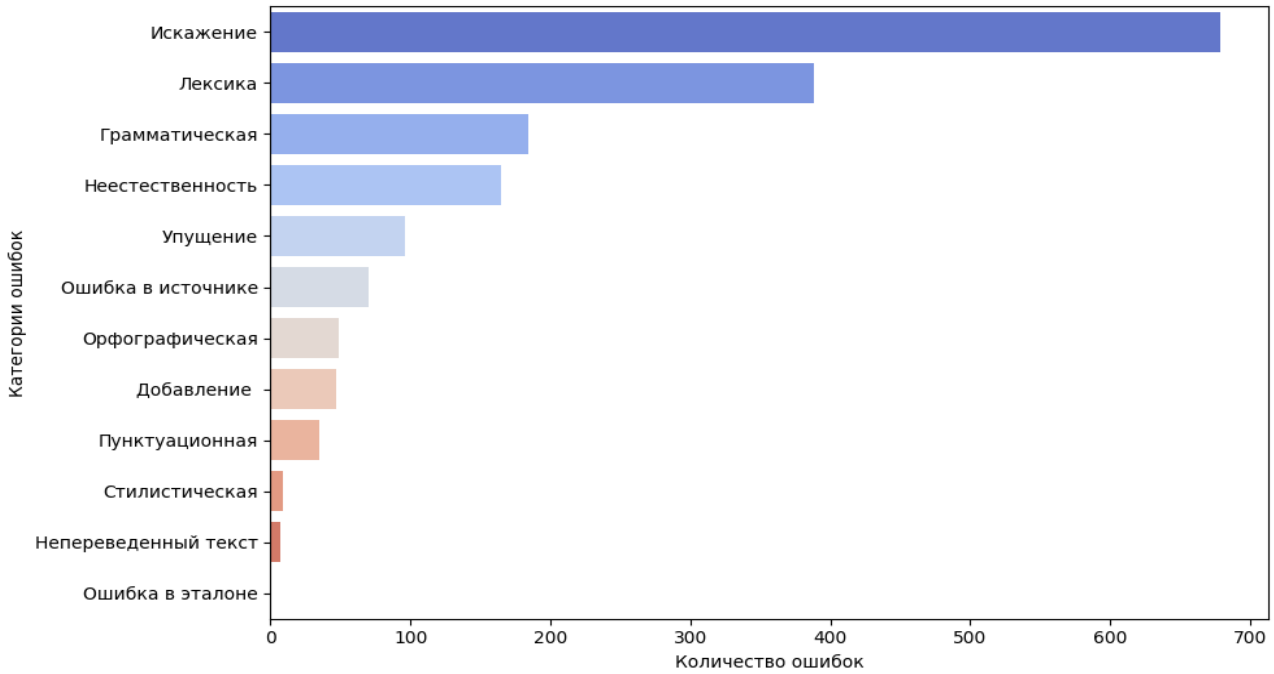

Рис. 9 Распределение ошибок по категориям (с мансийского на русский)

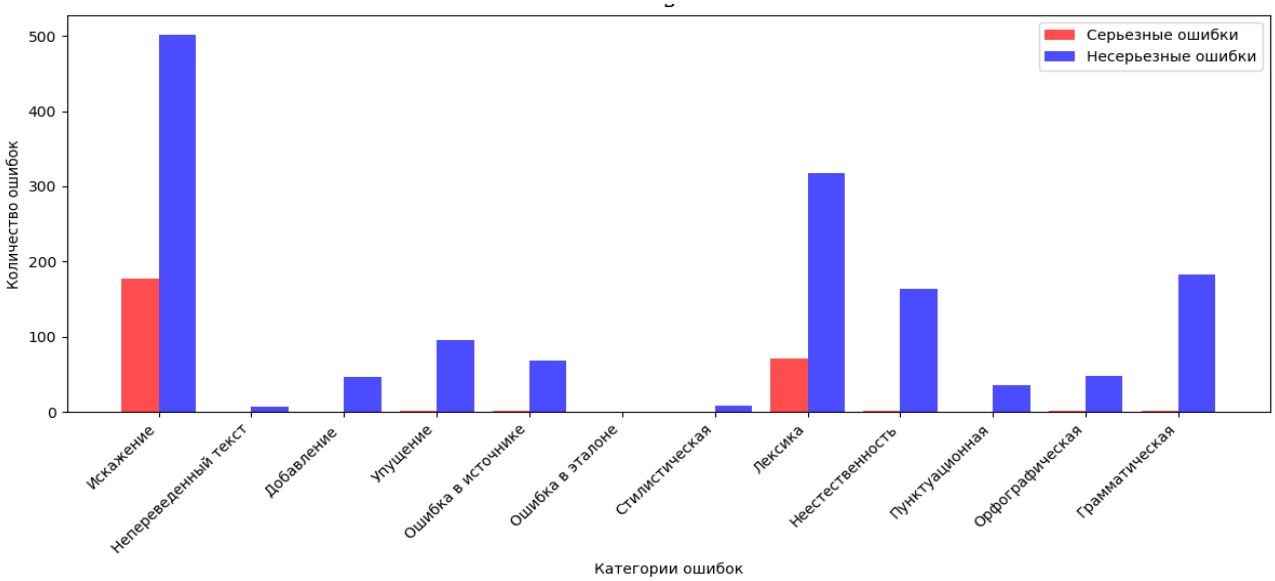

Рис. 10 Распределение ошибок по категориям и уровням серьезности (с мансийского на русский)

Исходя из всех этих графиков можно сделать следующие выводы. Перевод с русского на мансийский язык оказался значительно более качественным, по сравнению с переводом с мансийского на русский язык. Это также было заметно и на автоматических метриках. Эксперт отметил, что перевод на русский часто содержит множество грамматических и орфографических ошибок. Перевод с русского на мансийский язык содержит множество ошибок, связанных с грамматикой и неестественностью перевода. Наиболее критичными являются ошибки, приводящие к искажению смысла. При переводах с мансийского на русский язык также наблюдается много искажений.

В ходе результатов экспертной оценки были выявлено множество отдельных типов ошибок. Для улучшения качества перевода были сформулированы следующие наиболее частотные проблемы и предложены решения для улучшения качества перевода, которые могут быть учтены при доработке модели.

Пассивные конструкции. В мансийском языке частотны конструкции с пассивным залогом. Структурно такие предложения схожи с русскими примерами вида «Дом строится рабочими», однако пассивный залог в мансийском языке значительно более распространен и используется с широким классом глаголов, с частью которых пассивный залог в русском языка невозможен. Например, русское предложение «К нам пришло войско» можно перевести на мансийский язык с помощью пассивного залога: “Мāн хōнтна ёхтувесув” (букв. «Мы были приведены войском»). В связи с этой разницей в употреблениях модель с трудом обрабатывает перевод пассивных конструкций с мансийского на русский, и при переводе в обоих направлениях нарушается синтаксис таких конструкций. Для улучшения перевода необходимо добавить множество предложений с пассивными конструкциями. Чем больше таких примеров будет, тем выше вероятность, что модель сможет обобщить их использование. Создание подходящих предложений для перевода русского на мансийский может быть сложным, однако с мансийским языком таких проблем не возникает. В связи с этим рекомендуется сначала создавать предложения на мансийском языке, а затем проверять, улучшится ли результат работы модели. Формирование банка предложений должно осуществляться экспертами.

Перевод омонимов и многозначных слов. В мансийском языке частотны омонимичные и многозначные слова, которые часто переводятся неверно. Для улучшения способности модели различать значения омонимов и многозначных слов в разных контекстах следует добавить в набор данных примеры предложений с такими словами. Необходимо также составить список омонимов и на их основе создавать предложения, в которых эти слова будут использоваться в нужном контексте. Сбором частотных омонимов и созданием предложений на их основе должны заниматься эксперты.

Фольклорные тексты. В корпусе предложений существенную часть составляют предложения из традиционных фольклорных текстов. В них часто используются специфическая фольклорная лексика и особые речевые формулы, которые теряют смысл при дословном переводе. При переводе фольклорных текстов, таких как песни, важно учитывать их контекст. Если дословный перевод затрудняет понимание, необходимо стараться передать основной смысл. Эксперты должны составить список устойчивых выражений, характерных для фольклорных текстов, которые требуют особого внимания при переводе.

Перевод парных объектов. В мансийском языке парные объекты (например, руки, ноги, сапоги) грамматически устроены иначе, чем в русском. В русском языке используется множественное число (две руки) или слово «пара» (пара сапог), в то время как в мансийском существительное в единственном числе уже обозначает пару (лāгыл – форма единственного числа со значением «две ноги»), а один из элементов пары обозначается не просто единственным числом, как в русском (нога), а конструкцией со словом «половина» (лāгыл пāл дословно переводится как «половина ноги», однако обозначает одну ногу). Для создания предложений, содержащих парные объекты, можно воспользоваться генерацией на русском языке с помощью языковых моделей (LLM) либо же эксперты могут подготовить список предложений с употреблением лексики, обозначающей парные объекты.

Порядок слов. Одной из частотных, хоть и не критичных ошибок является порядок слов. В мансийском языке часто при переводе наблюдается неестественный порядок слов, который в некоторых случаях также является ошибочным. В частности, при переводе на мансийский язык часто возникает проблема с неправильным расположением частицы отрицания.

Также были сформулированы рекомендации для дальнейшей работы экспертов и переводчиков, которые создают корпус параллельных предложений:

- необходимо выработать единое решение по мансийской орфографии для разрабатываемой модели, которое будет отражать предпочтения сообщества, но также позволять унифицированное представление текста на мансийском языке в корпусе;
- необходимо учитывать наличие устойчивых выражений и осуществлять смысловой, а не дословный перевод;

- если дословный перевод искажает или затрудняет передачу смысла оригинального предложения, следует сосредоточиться на передаче основного смысла текста;
- при пополнении корпуса необходимо использовать целые предложения, а не отдельные фрагменты, которые невозможно верно интерпретировать без учета контекста.

Заключение

В рамках данной работы была обучена модель нейронного машинного перевода для языковой пары русский–мансийский. В процессе реализации использовались библиотеки PyTorch и Transformers, в ходе обучения моделей лучшей моделью оказалась NLLB-200-3.3B для перевода с русского на мансийский и обратно.

Для оценки качества перевода применялись метрики BLEU и chrF, по результатам которых модель достигла показателей BLEU 27 % и chrF 57 % для переводов с русского на мансийский, и 25.3 % и chrF 50.3 % с мансийского на русский. Эти результаты для данных метрик являются довольно хорошими. Однако основное качество модели было подтверждено экспертной оценкой, в ходе которой были выявлены её сильные и слабые стороны. Для экспертной оценки была адаптирована многомерная метрика качества машинного перевода, исходя из человеческих ресурсов и задач, которая позволяет отразить ошибки, совершаемые машинным переводом. С помощью экспертной оценки были определены направления для улучшения модели, а также разработаны планы по расширению и совершенствованию корпуса мансийского языка.

Результаты показали эффективность применения трансформерных моделей для задачи машинного перевода. Разработанная модель продемонстрировала неплохое качество перевода на данном этапе и может быть использована в практических приложениях. Полученные результаты и накопленный опыт будут полезны для дальнейших исследований и разработок в области машинного перевода.

Благодарности

Выражаем благодарность носителям мансийского языка, принимающим участие в проекте в качестве переводчиков и экспертов, за сбор корпуса параллельных предложений для мансийского языка. Их работа является самой важной в создании данных для обучения моделей для машинного перевода. Поэтому мы выражаем благодарность коллективу газеты «Лӯимā сэ̄рипос» за переводы предложений на мансийский язык, а также другим переводчикам, которые принимали активное участие в проекте. Реализация данного проекта была возможна благодаря поддержке в рамках Государственного задания, что позволило успешно собрать корпус и заложить основу для дальнейших исследований.

Список литературы / References

[1] Transfer Learning for Low-Resource Neural Machine Translation. Available at: https://arxiv.org/pdf/1604.02201

[2] Trivial Transfer Learning for Low-Resource Neural Machine Translation. Available at: https://aclanthology.org/W18-6325.pdf

[3] Multilingual Denoising Pre-training for Neural Machine Translation. Available at: https://arxiv.org/pdf/2001.08210

[4] Extending the Subwording Model of Multilingual Pretrained Models for New Languages. Available at: https://arxiv.org/pdf/2211.15965

[5] Low-Resource Multilingual Neural Translation Using Linguistic Feature-based Relevance Mechanisms. Available at: https://dl.acm.org/doi/10.1145/3594631

[6] When and Why are Pre-Trained Word Embeddings Useful for Neural Machine Translation? Available at: https://arxiv.org/pdf/1804.06323

[7] Pre-Training Multilingual Neural Machine Translation by Leveraging Alignment Information. Available at: https://aclanthology.org/2020.emnlp-main.210.pdf

[8] No Language Left Behind: Scaling Human-Centered Machine Translation. Available at: https://arxiv.org/pdf/2207.04672

[9] MADLAD-400: A Multilingual and Document-Level Large Audited Dataset. Available at: https://arxiv.org/pdf/2309.04662

[10] BLEU: a Method for Automatic Evaluation of Machine Translation. Available at: https://aclanthology.org/P02-1040.pdf

[11] METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. Available at: https://aclanthology.org/W05-0909.pdf

[12] A Study of Translation Edit Rate with Targeted Human Annotation. Available at: https://aclanthology.org/2006.amta-papers.25.pdf

[13] CHRF: character n-gram F-score for automatic MT evaluation. Available at: https://aclanthology.org/W15-3049.pdf

[14] BERTSCORE: EVALUATING TEXT GENERATION WITH BERT. Available at: https://arxiv.org/pdf/1904.09675

[15] COMET: A Neural Framework for MT Evaluation. Available at: https://arxiv.org/pdf/2009.09025

[16] MQM (Multidimensional Quality Metrics). Available at: https://themqm.org/

[17] The MQM-Full Master File Instructions. Available at: https://themqm.org/wp-content/uploads/2024/03/MQM-Full-Master-Instructions_2024_01_30.pdf

[18] Multi-Dimensional Machine Translation Evaluation: Model Evaluation and Resource for Korean. Available at: https://arxiv.org/html/2403.12666v1

[19] Translation Quality Assessment: MQM (Multidimensional Quality Metrics). Available at: https://sites.miis.edu/runyul/2018/03/04/translation-quality-assessment-mqm-multidimensional-quality-metrics/

[20] Expert-based Human Evaluations for the Submissions of WMT 2020, WMT 2021, WMT 2022 and WMT 2023. Available at: https://github.com/google/wmt-mqm-human-evaluation/blob/main/README.md

[21] Results of WMT23 Metrics Shared Task: Metrics might be Guilty, but References are not Innocent. Available at: https://www2.statmt.org/wmt23/pdf/2023.wmt-1.51.pdf

[22] Summary of the tokenizers. Available at: https://huggingface.co/docs/transformers/tokenizer_summary

[23] How to fine-tune a NLLB-200 model for translating a new language. Available at: https://cointegrated.medium.com/how-to-fine-tune-a-nllb-200-model-for-translating-a-new-language-a37fc706b865

Поступила в редакцию 20 декабря 2024 г.

МЕТАДАННЫЕ / METADATA

Title: Development of neural machine translation model for the Mansi language.

Abstract: The paper presents a description of the process of training a transformer-based neural network for solving the machine translation task for the Mansi language (Ob-Ugric < Finno-Ugric < Uralic), which is a low-resource language. The aim of the study is to conduct experiments comparing the results of fine-tuning multilingual models for the language pair: Russian and Mansi. The paper provides an overview of modern machine translation methods and neural network architectures, including transformer networks. As part of the study, neural networks were fine-tuned using the PyTorch and Transformers libraries. Translation quality was evaluated using BLEU and chrF metrics. The best result was achieved with the NLLB-200-3.3B model, which attained BLEU 27% and chrF 57% for translation from Russian to Mansi. Additional experiments and analyses were conducted to identify the strengths and weaknesses of the methods through expert evaluation. The study demonstrates the effectiveness of applying transformer models to the machine translation task and can be used in practical applications.

Key words: low-resource languages; machine translation; Finno-Ugric languages; Mansi language.

Язык статьи / Language: Русский / Russian.

Об авторах / About the authors:

НЕГМАТУЛОЕВ Одилжон Олимжонович
Югорский государственный университет, Россия.
Аспирант инженерной школы цифровых технологий (ЮГУ).
E-mail: odilzhon100@gmail.com
ORCID:https://orcid.org/0009-0006-2165-8962

ЖОРНИК Дарья Олеговна
Институт языкознания РАН, Россия.
Научный сотрудник института.
E-mail: daria.zhornik@yandex.ru
ORCID: https://orcid.org/0000-0002-6463-2547

МЕЛЬНИКОВ Андрей Витальевич
Югорский НИИ информационных технологий, Россия.
Директор института.
E-mail: melnikovav@uriit.ru
ORCID: https://orcid.org/0000-0002-1073-7108

NEGMATULOEV Odilzhon Olimzhonovich
Ugra State University, Russia.
PhD student in the School of Digital Engineering.
E-mail: odilzhon100@gmail.com
ORCID:https://orcid.org/0009-0006-2165-8962

ZHORNIK Darya Olegovna
Institute of Linguistics of the RAS, Russia.
Research Fellow.
E-mail: daria.zhornik@yandex.ru
ORCID: https://orcid.org/0000-0002-6463-2547

MELNIKOV Andrey Vitalevich
Ugra Research Institute of Information Technologies, Russia
Director of the Research Institute.
E-mail: melnikovav@uriit.ru
ORCID: https://orcid.org/0000-0002-1073-7108