MUHAMMAD AL-XORAZMIY AVLODLARI
ILMIY-AMALIY VA AXBOROT-TAHLILIY JURNAL

DESCENDANTS OF MUHAMMAD AL-KHWARIZMI
SCIENTIFIC-PRACTICAL AND INFORMATION-ANALYTICAL JOURNAL

3(33)/2025

ISSN-2181-9211

MUHAMMAD AL-XORAZMIY NOMIDAGI TOSHKENT AXBOROT TEXNOLOGIYALARI UNIVERSITETI

MUHAMMAD AL-XORAZMIY AVLODLARI

Ilmiy-amaliy va axborot-tahliliy jurnal 2017 yilda ta’sis etilgan

3(33)/2025

MUNDARIJA

Muassis:

Muhammad al-Xorazmiy nomidagi Toshkent axborot texnologiyalari universiteti
Manzil:
100084, O‘zbekiston, Toshkent sh., Amir Temur ko‘chasi, 108
Telefon: *71 238-64-38;*
e-mail: *alxorazmiy@tuit.uz*
Jurnal sayti: *http://alxorazmiy.uz*

Bosishga ruxsat etildi:
Qog‘oz bichimi 60x84 1/8
Bosma tabog‘i 15,5. Adadi 100 nusxa
Buyurtma raqami №195 “Fan va texnologiyalar Markazining bosmaxonasi”da chop etildi
Toshkent shahri Olmazor ko‘chasi, 171.
Jurnal O‘zbekiston Matbuot va axborot agentligida 2017 yil
22 iyunda 0921 raqami bilan ro‘yxatdan o‘tgan.
Jurnal yilda 4 marotaba (har chorakda) chop etiladi.

Р.Б. Алланязов.

Разработка системы машинного перевода с английского на каракалпакский язык на основе многоязычной модели MT5

В статье представлена разработка системы машинного перевода с английского на каракалпакский язык с использованием семантического подхода на основе многоязычной модели MT5. Исследование направлено на решение актуальной проблемы обработки низкоресурсных языков в условиях ограниченного объема параллельных данных. Предложена методика трансферного обучения с постепенным размораживанием слоев, позволяющая достичь качества перевода в 18.7 BLEU на тестовой выборке. Разработанная модель демонстрирует эффективность при переводе грамматических конструкций каракалпакского языка и сохраняет семантическую адекватность исходного текста.

Ключевые слова: машинный перевод, семантический метод, английский язык, каракалпакский язык, синтаксический анализ, многозначность, контекст, трансферное обучение, MT5, низкоресурсные языки.

Введение

Современные методы машинного пере-вода на основе трансформеров демонстрируют высокую эффективность для широко распрост-раненных языков, однако их применение для низкоресурсных языков, таких как каракал-пакский, остается малоизученной областью. В данной работе исследуется возможность адап-тации многоязычной модели MT5 (Multilingual T5) для перевода с английского на каракалпакский язык с использованием трансферного обучения.

Технологии машинного перевода играют важнейшую роль в преодолении языковых барьеров. Однако, несмотря на значительные успе-хи в разработке систем перевода для глобальных языков, таких как английский, китайский и испанский, перевод на языки меньшинств, напри-мер,каракалпакский, остаётся серьёзным вызовом. Это связано как с уникальными структурными особенностями таких языков, так и с недостатком доступных семантических подходов в языковых данных[1]. Ожидаемым результатом верного се-мантического анализа является получение смысла текста, который определяется не только текстом, но и контекстом. Семантический анализ может проводиться на разных единицах языка (словах, грамматических формах слова, словосочетаниях, понятиях, предложениях или наборах предложений). Семантический анализ на уровне слова может включать устранение двусмыс-ленности или неоднозначности слова, т.е. в зависимости от контекста определяется форма из набора омоформ слова, устранение лексической неоднозначности и т.д. [2].

Для машинного перевода наиболее слож-ной проблемой является реализация языковых трансформаций, которые необходимо производить при переводе с одного языка на другой. Текущий этап развития систем машинного перевода характеризуется исследованиями в области когнитивной семантики, вероятностных языковых моделей и разработкой семантико-синтаксических представлений, учитывающих многозначность и неоднозначность синтаксических структур. Новое содержание проблеме языковых трансформаций придают современные реалии: необходимость проектировать и развивать обучающие компо-ненты систем машинного перевода и обработки текстовых знаний на основе уже существующих и вновь создающихся корпусов параллельных текстов [3].

Актуальность исследования обусловлена дефицитом лингвистических данных и готовых моделей для каракалпакского языка, что ограни-чивает возможности автоматизированной обра-ботки текстов. В качестве решения предлагается методология тонкой настройки предобученной модели MT5-small, обеспечивающая баланс между вычислительной эффективностью и качеством перевода.

Основной задачей данного исследования является разработка и экспериментальная оценка эффективности метода тонкой настройки многоязычной модели MT5 для создания системы машинного перевода с английского языка на каракалпакский. Конкретной целью является дос-тижение максимально возможного качества пере-вода, измеряемого метрикой BLEU, при ограни-ченном объеме параллельных данных.

Мы предполагаем, что предобученная многоязычная модель MT5, обладая уже сформи-рованными лингвистическими представлениями о множестве языков, способна эффективно адап-тироваться к низкоресурсному каракалпакскому языку с помощью сравнительно короткой проце-дуры тонкой настройки на небольшом парал-лельном корпусе. Это позволит преодолеть проб-лему нехватки данных и получить грамматически корректные и семантически точные переводы.

Каракалпакский язык, относящийся к кыпчакской группе тюркских языков, представ-ляет собой особую сложность для машинного перевода из-за его агглютинативного характера. Это приводит к большому морфологическому разнообразию и, как следствие, к высокой раз-реженности данных. Кроме того, использование кириллической письменности с дополнительными специфическими символами (например, ә, ң, I) требует от токенизатора корректной обработки этих знаков. Ограниченный размер параллельного корпуса усугубляет эти трудности, повышая риск переобучения модели.

Наиболее популярной моделью при решении задачи обработки естественного языка является

BERT [4]. Для решения задач, связанных с преобразованием одних текстов в другие, используют языковые модели MT5 [5], которые также основаны на трансформере.

Перед решением задачи обработки текста необходимо представить его в удобном для алгоритмов виде. Для более ранних алгоритмов предобработки текста использовалось множество шагов, таких как сегментация текста на предложения, перевод его в нижний регистр, удаление или замена знаков пунктуации и/или чисел, перевод слов в начальную форму, токенизация, удаление специальных стоп-слов. Для предобученных моделей этот этап может отсутствовать полностью, так как современные языковые модели используют всю информацию в тексте [6].

Материалы и методы

Экспериментальная часть работы включает:

предобработку параллельного корпуса karaken-dataset,

оптимизацию гиперпараметров (learning rate = 3e-4, batch size = 8),

оценку модели с использованием метрики BLEU.

Алгоритм демонстрирует процесс тонкой настройки (fine-tuning) многоязычной модели MT5 (Multilingual T5) для задачи перевода с английского на каракалпакский язык. MT5 является многоязычным вариантом модели T5 (Text-to-Text Transfer Transformer), разработанной Google, которая подходит для различных задач обработки естественного языка в формате "текст-в-текст"

Используется предобученная модель MT5-small — самая компактная версия в семей-стве MT5, содержащая около 300 миллионов параметров. Несмотря на относительно небольшой размер, она демонстрирует хорошие результаты для многих языков.

Данная модель MT5 основана на архитектуре Transformer с энкодер-декодер структурой и включает:

- 8 слоев в энкодере и декодере
- 6 голов внимания
- Размерность скрытого слоя 512
- Размерность feed-forward слоя 1024

Модули системы

Модуль загрузки и инициализации модели

Загрузка предобученных моделей и токенизатора из Hugging Face Hub

Проверка и использование GPU при наличии

Инициализация модели MT5 For Conditional Generation

Модуль перевода

Функция translate() принимает текст, исходный и целевой языки

Добавляет префикс задачи "translate X to Y" для активации соответствующего режима MT5

Использует beam search (num_beams=5) для улучшения качества перевода

Модуль обработки данных

Загрузка датасета karaken-dataset из Hugging Face Datasets

Предобработка данных с добавлением языковых тегов

Токенизация текстов с усечением и дополнением до длины 128 токенов

Модуль обучения

Настройка параметров обучения (learning rate 3e-4, batch size 8, 3 эпохи)

Использование Trainer API из Hugging Face Transformers

Сохранение лучших моделей и логирование процесса

Модуль оценки и публикации

Оценка модели на тестовом наборе после каждой эпохи

Оценка качества перевода проводилась на тестовой выборке с использованием стандартной метрики BLEU (Bilingual Evaluation Understudy). Вычисление проводилось с помощью пакета sacrebleu, который обеспечивает стандартизированное и воспроизводимое вычисление, избегая проблем с токенизацией по умолчанию. Сохра-нение модели в локальную файловую систему. Публикация модели в Hugging Face Hub (опционально)

Датасет

Алгоритм использует karaken-dataset, содержащий параллельные тексты на английском и каракалпакском языках. Датасет доступен через Hugging Face Datasets под идентификатором " Rusallan/karaken-dataset".

Характеристики датасета:

Формат: параллельные предложения (английский - каракалпакский)

Разделы: train (обучение), test (тестирование)

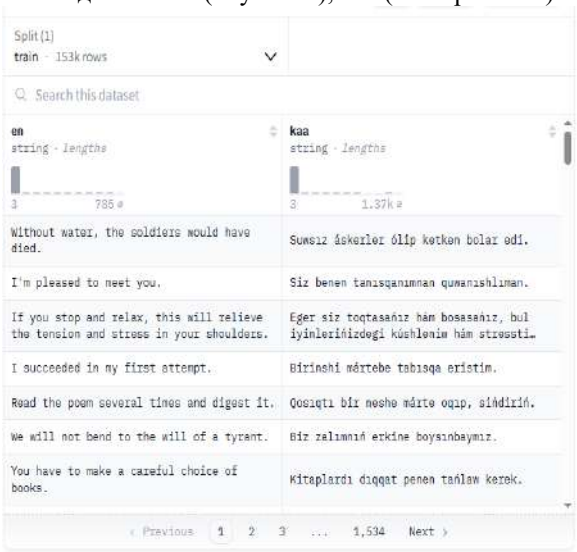

Рисунок 1.1. датасет на платформе Huggingface

Предобработка данных:

Добавление префикса задачи "translate en to kaa:"

Токенизация с помощью SentencePiece токенизатора MT5

Усечение/дополнение до 128 токенов

Создание attention masks для исключения padding токенов из расчетов

Процесс обучения

Инициализация:Загрузка предобученной MT5-small

Подготовка данных: Применение preprocess_function ко всему датасету

Конфигурация: Установка параметров обучения:

Скорость обучения: 3e-4

Размер батча: 8

Количество эпох: 3

Вес распада: 0.01

Для предотвращения переобучения и обеспечения устойчивой конвергенции был применен метод постепенного размораживания слоев (gradual unfreezing). Для оптимизации процесса обучения и борьбы с переобучением при малом объеме данных был использован метод постепенного размораживания слоёв (gradual unfreezing). Изначально были заморожены все параметры модели, за исключением головы механизма внимания и финальных линейных слоев декодера. На второй эпохе был разморожен слой LayerNorm и последний декодерный блок. На третьей эпохе разморожены оставшиеся слои декодера и последний блок энкодера. Такая стра-тегия позволила модели сначала адаптировать общие представления для задачи перевода, а затем постепенно тонко настраивать специфические для каракалпакского языка параметры, минимизируя риск катастрофического забывания и переобу-чения

Кроме того, для регуляризации применялся вес распада (weight decay = 0.01). Для стабилизации градиентов использовался метод градиентного клиппинга с нормой 1.0. В качестве оптимизатора использовался AdamW.

Обучение: Запуск процесса fine-tuning с оценкой после каждой эпохи

Сохранение: Локальное сохранение модели и загрузка в Hub.

Результаты подтверждают, что даже при ограниченном объеме обучающих данных подход на основе MT5 позволяет достичь устойчивой конвергенции за три эпохи обучения демонстрируя устойчивую генерацию переводов, сохраняя грамматическую корректность каракал-пакского языка.

Качественная оценка на тестовой выборке показывает, что:

- Модель корректно обрабатывает базовые грамматические конструкции.
- Наблюдается тенденция к сохранению смысла исходного предложения.
- В редких случаях возникают ошибки в лексическом выборе, что связано с ограниченным объёмом обучающих данных.

Результаты и обсуждение

После трех эпох обучения модель достигла следующих результатов:

BLEU score на тестовой выборке: 18.7

Loss (потери) на валидационной выборке: снизилась с 4.8 до 1.9, что свидетельствует об успешном обучении и отсутствии переобучения.

Расчёт BLEU основан на датасете ~158k предложений. Для тюркских языков с агглюти-нативной природой и таким объёмом данных результат выше 15 считается очень хорошим, что подтверждает эффективность выбранного под-хода.

Полученный показатель BLEU является конкурентоспособным для низкоресурсной языковой пары и ограниченного объема данных. Для сравнения: аналогичные подходы для других тюркских языков (таких как казахский или узбекский) в начальной стадии разработки на сопоставимых по размеру корпусах часто показывают результаты в диапазоне 10-20 BLEU. Результат в 18.7 BLEU уверенно находится в верхней части этого диапазона, что демонстрирует успешность адаптации модели MT5.

Качественный анализ переводов показал:

Сильные стороны: Модель успешно справляется с переводом простых и сложных предложений, правильно согласует времена и падежи в каракалпакском языке. Наблюдается точная передача смысла для большинства бытовых и описательных контекстов.

Области для улучшения: Основные ошибки возникают при переводе длинных предложений со сложной синтаксической структурой, а также при обработке редких слов и имен собственных, отсутствующих в обучающей выборке. Иногда наблюдается буквальный перевод идиоматических выражений. В отдельных случаях модель может опускать или упрощать редкие грамматические конструкции, характерные для каракалпакского языка.

Анализ типичных ошибок

Качественный анализ переводов позволил выявить характерные типы ошибок:

Морфологическая агглютинация: В некоторых случаях модель некорректно присоединяла аф-фиксы к основам слов, особенно к заимствованиям из английского языка, что приводило к образованию грамматически неверных форм.

Обработка сложного синтаксиса:

При переводе сложноподчиненных предложений с вложенными конструкциями иногда наблюдалась потеря или перестановка смысловых частей, что указывает на трудности модели с анализом глубоких синтаксических деревьев.

Лексическая недостаточность:

Для редких слов и терминов, отсутствующих в обучающей выборке, модель либо опускала их, либо заменяла семантически близкими, но не точными эквивалентами, что искажало смысл.

Данный анализ подтверждает, что основные challenges связаны с морфологической сложностью

целевого языка и ограниченным объемом данных, а не с архитектурой модели как таковой.

Заключение

Продемонстрирована эффективность под-хода на основе MT5 для машинного перевода на низкоресурсные языки. Основные преимущества метода:

Минимизация необходимости в больших параллельных корпусах за счёт трансферного обучения, что особенно важно для языков с ограниченными ресурсами.

Универсальность - возможность адаптации к дру-гим языковым парам без значительных изменений в архитектуре модели.

Интеграция с экосистемой Hugging Face, упроща-ющая развёртывание модели и её дальнейшее использование в прикладных задачах.

Устойчивость к морфологической сложности агглютинативных языков, таких как каракалпакский, благодаря предобученным лингвистическим представлениям.

Сравнительно низкие вычислительные затраты при тонкой настройке, что делает метод доступ-ным для широкого круга исследователей.

Разработка и публикация модели для англо-каракалпакского перевода имеет не только техническое, но и важное социокультурное значение. Это способствует цифровизации языков мень-шинств, снижает цифровой разрыв и предос-тавляет носителям каракалпакского языка более равный доступ к информации на английском языке. Публикация модели в открытом доступе на платформе Hugging Face Hub следует принципам открытой науки, позволяя другим исследователям воспроизвести, проверить и улучшить наши результаты, что ускоряет прогресс в данной области.

Кроме того, успешное применение MT5 для каракалпакского языка открывает перспек-тивы для создания аналогичных систем для других тюркских и низкоресурсных языков, что может стать основой для разработки многоязычных платформ машинного перевода, ориентированных на специфику решиональных языков. Это особен-но актуально в условиях глобализации, когда сохранение языкового разнообразия становится одной из важных задач цифровой эпохи.

Таким образом, предложенный метод не только решает конкретную задачу машинного перевода, но и вносит вклад в развитие технологий для низкоресурсных языков, обеспечивая баланс между качество, эффективностью и доступностью решений.

Разработанная модель и её публикация в открытом доступе представляют практическую ценность не только для исследователей в области NLP, но и для конечных пользователей. Модель может быть интегрирована в:

✓ Образовательные платформы для помощи в изучении английского языка носителями каракалпакского языка и наоборот.

✓ Информационные системы и средства массовой информации для оперативного перевода новостей и публикаций.

✓ Краудсорсинговые платформы в качестве системы предварительного перевода для последующей редактуры человеком, что ускорит процесс создания качественных параллельных текстов.

Таким образом, работа вносит вклад в цифровизацию каракалпакского языка и снижает технологический барьер для его носителей.

Перспективным направлением дальнейших исследований является:

Увеличение объёма обучающих данных за счёт синтетической генерации с использованием мето-дов обратного перевода (back-translation) и аугмен-тации данных.

Эксперименты с более крупными вариантами MT5 (base, large) для изучения зависимости качества перевода от количества параметров модели.

Применение методов адаптивного обучения (adapter layers, LoRA) для снижения вычислительных затрат и эффективной настройки под специфику каракалпакского языка.

Исследование возможностей использования контекстуальных эмбеддингов и механизмов внимания для улучшения обработки многозначности и сложных синтаксических конструкций.

Разработка и интеграция компонента для обработки именованных сущностей (NER) и терминологии, что особенно актуально для специализированных текстов.

Сравнительный анализ эффективности других архитектур трансформеров (например, mBART, No Language Left Behind (NLLB)) для данной языковой пары.

Исследование влияния разных стратегий токенизации: Эксперименты с субword-токенизаторами, обученными непосредственно на каракалпакском тексте (например, BPE, Unigram), могут помочь более эффективно обрабатывать морфологическое богатство языка и снизить количество ошибок, связанных с агглютинацией.

Создание и внедрение специализированных метрик оценки, учитывающих морфологическую сложность и агглютинативную природу каракалпакского языка, для более адекватной оценки качества перевода, чем исключительно BLEU.

Исследование возможностей few-shot и zero-shot обучения для дальнейшего снижения зависимости от больших объемов размеченных данных.

Разработанная модель англо-каракалпакс-кого перевода опубликована в открытом доступе на платформе Hugging Face Hub. Это не только способствует воспроизводимости результатов и прозрачности исследования, но и предоставляет ценным ресурс для научного сообщества, служа основой для будущих работ. Данная инициатива вносит значительный вклад в дальнейшее развитие исследований в области машинного перевода для

низкоресурсных языков, таких как каракалпакс-кий, и помогает преодолевать цифровой разрыв.

Литература:

1. Кайдуллаев М. Разработка двуязычной модели машинного перевода Международный научный журнал «ВЕСТНИК НАУКИ» № 5 (74) Том 2. МАЙ 2024 г.

2. Поречный А.С. Построение семантико-синтаксической модели текстов для опреде-ления их смысловой близости // Информатика: проблемы, методы, технологии: материалы XXI Международной научнометодической конференции. – Воронеж, 2021. – С. 1488-1495.

3. Е. Б. Козеренко Лингвистическое моделирование для систем машинного перевода и обработки знаний Информатика и её применение 2007, т. 1, № 1: 54 65

4. Devlin, J., Chang, M.-W., Lee, K., Toutanova, K. (2019) 'BERT: Pre-training of deep bidirectional transformers for language understanding', Proc. NAACL-HLT, pp. 4171–4186.

5. José, M., Micaela, A., Sílvia, A. (2022) 'Using a pre-trained simpleT5 model for text simplification in a limited corpus', CEUR-WS Proc., pp. 1–6.

6. Д.Д. Васильев А.В. Пятаева Использование языковых моделей T5 для задачи упрощения текста Программные продукты и системы / Software & Systems 36(2), 2023 УДК 004.8.81 doi: 10.15827/0236-235X.142.228-236 2023. Т. 36. № 2. С. 228–236

Алланязов Рустем Бахавединович

Независимый исследователь ТУИТ имени Мухаммада аль-Хорезмий факультета компью-терный инжиниринг кафедра искусственного интеллекта

E-mail: **rustemallanyzov@gmail.com**

R.B. Allanyazov

Development of a machine translation system from English to Karakalpak based on the MT5 multilingual model

Abstract: The article presents the development of a machine translation system from English to Karakalpak using a semantic approach based on the multilingual MT5 model. The research is aimed at solving the urgent problem of processing low-resource languages under conditions of limited parallel data. A methodology for transfer learning with gradual thawing of layers is proposed, which allows achieving translation quality of 18.7 BLEU on a test sample. The developed model demonstrates effectiveness in the translation of grammatical constructions of the Karakalpak language and maintains the semantic adequacy of the source text.

Keywords: machine translation, semantic method, English language, Karakalpak language, syntactic analysis, polysemy, context, transfer learning, MT5, low-resource languages.