DOI: 10.33184/dokbsu-2026.2.20

Архитектура Seq2Seq для малоресурсных языков: методы преодоления дефицита данных

Р. Г. Мифтахова*, Н. М. Мыльников

Уфимский университет науки и технологий
Россия, Республика Башкортостан, 450076 г. Уфа, ул. Заки Валиди, 32.

**Email: miftahovar@yandex.ru*

Представленное исследование содержит подходы к модификации моделей Seq2Seq для сценариев обработки малоресурсных языков. Рассматриваются ключевые методы преодоления «проклятия размерности» и разреженности данных: алгоритмы субсимвольной токенизации (BPE, SentencePiece); методы многоязыкового кросс-языкового трансфера (multilingual transfer learning) с использованием высокоресурсных языков-доноров; а также методы синтетической аугментации данных (back-translation). Отдельное внимание уделяется современным методам параметрически эффективного дообучения (PEFT). В работе представлен сравнительный анализ продуктивности описанных подходов, стратифицированный в соответствии с размером доступных параллельных данных.

Ключевые слова: машинный перевод, обработка естественного языка, Sequence-to-Sequence, малоресурсные языки, трансферное обучение, обратный перевод, LoRA.

В последнее десятилетие технологии обработки естественного языка (Natural Language Processing, NLP) переживают беспрецедентный скачок развития. Системы нейронного машинного перевода (Neural Machine Translation, NMT), основанные на архитектурах Sequence-to-Sequence (Seq2Seq), достигли и в некоторых случаях превзошли уровень человеческого качества перевода. Однако этот успех крайне неравномерен в мировом языковом многообразии. Подавляющее большинство исследований и передовых метрик опирается на высокоресурсные языки (английский, испанский, китайский, русский), для которых существуют параллельные корпусы объемом в десятки миллионов предложений.

При этом из более чем 7 000 языков мира свыше 90% классифицируются как малоресурсные (low-resource languages). Для них характерно отсутствие масштабных оцифрованных текстовых баз, электронных словарей и размеченных параллельных датасетов. Применение классического сквозного обучения (end-to-end) глубоких нейронных сетей к таким языкам приводит к проблеме сильного переобучения на обучающей выборке и неспособности модели к обобщению (generalization) [1, с. 6000]. Целью данной

статьи является систематизация и детальный анализ передовых математических и архитектурных методов оптимизации Seq2Seq-моделей в условиях жесткого дефицита обучающих данных.

Проблема переобучения нейронных сетей в условиях дефицита данных носит фундаментальный характер и обусловлена высокой емкостью моделей Seq2Seq. При обучении трансформеров на выборках размером менее 10 000 параллельных предложений, ландшафт функции потерь (loss landscape) становится крайне неравномерным. Модель достигает быстрой минимизации эмпирического риска через примитивное запоминание обучающих примеров, что ведет к существенному разрыву в обобщении. Градиенты при обратном распространении начинают кодировать шум и артефакты ограниченной выборки, а не реальные языковые паттерны. На данных вне распределения наблюдается высокая энтропия ответов и склонность к галлюцинациям. Для малоресурсных языков проблема усугубляется дефицитом качественной предобработки, что повышает уровень шума в корпусах и блокирует сходимость при обучении с нуля.

Рассмотрим математические основы архитектуры Seq2Seq и механизм внимания. Классическая задача машинного перевода в парадигме Seq2Seq формулируется как нахождение условной вероятности целевой последовательности $Y = (y_1, y_2, ..., y_n)$ при заданной входной последовательности $X = (x_1, x_2, ..., x_n)$. Модель обучается авторегрессионно, предсказывая следующий токен на основе контекста исходного предложения и уже сгенерированных токенов.

$$P(Y|X) = \Pi P(y_t \mid y_{<t}, X; \theta),$$

где θ – оптимизируемые параметры нейронной сети. В современной практике дефакто стандартом является архитектура Transformer, отказавшаяся от рекуррентных связей в пользу механизма скалярного внутреннего внимания (Scaled Dot-Product Attention) [1, с. 6001]. Функция реализуется формулой:

$$(Q, K, V) = \text{softmax}(Q K^T / \sqrt{d_k}) V,$$

в которой параметр d_k обозначает размерность пространства ключей. Механизм дает модели возможности для установления глобальных зависимостей между токенами, однако сильно зависит от размера обучающего корпуса. Следовательно, архитектура, отлично работающая на миллиардных датасетах, требует пересмотра инициализации и оптимизации для малоресурсной среды.

При работе с малоресурсными языками ключевым становится также выбор оптимальной токенизации. Субсимвольная позволяет бороться с богатой морфологией. Критическим барьером для малоресурсных языков (в частности, агглютинативных) является

обилие уникальных словоформ, приводящее к высокому проценту внесловарных токенов (Out-of-Vocabulary, OOV). Решением выступает алгоритм кодирования пар байтов (Byte-Pair Encoding, BPE) [5, с. 1716]. Вместо оперирования целыми словами, модель обучается на уровне подслов и морфем. Процесс BPE инициализируется словарем символов, после чего итеративно объединяются наиболее частотные смежные пары токенов. Это радикально уменьшает количество параметров на слое эмбеддингов.

Эффективность алгоритмов BPE и SentencePiece особенно ярко проявляется при работе с языками агглютинативного и полисинтетического типов, для которых характерен экспоненциальный рост уникальных словоформ. В таких языках одно слово может кодировать сложную семантико-синтаксическую конструкцию, включающую корень и множественные аффиксы. Ограниченный размер словаря, применяемый в классических методах, неизбежно приводит к потере критически важной информации при замене редких словоформ токеном [UNK]. Субсимвольная токенизация помогает решить данную проблему. Данный алгоритм разделяет сложные словоформы на привычные для системы частотные морфемы, позволяя нейросети формировать семантику незнакомого слова из известных подсловных единиц (subword units). Данное действие помогает экономить вычислительные ресурсы путем сжатия матрицы встраивания (embedding matrix) и формирует более плотное и непрерывное латентное пространство (latent space), группируя семантически близкие морфемы в локальных кластерах.

Настоящей проблемой может стать обучение нейронной сети в условиях малоресурсного языка. При наличии выборки размером менее 10000 параллельных предложений обучение сети «с нуля» математически нецелесообразно. Парадигма трансферного обучения решает эту проблему за счет переноса знаний из высокоресурсных языков [2, с. 3]. Современные модели, такие как mBART (Multilingual BART) [7, с. 728], предварительно обучаются на гигантских моноязычных корпусах десятков языков задаче реконструкции зашумленного текста. При дообучении на параллельной паре «малоресурсный ↔ высокоресурсный» модель адаптирует уже сформированные богатые языковые представления.

Фундаментальная гипотеза, лежащая в основе трансферного обучения (multilingual transfer learning), заключается в существовании универсального языкового пространства репрезентаций (shared semantic space). В процессе длительного предобучения (pre-training) на смешанном корпусе из десятков языков, модель mBART вынуждена кодировать грамматические и семантические концепты независимо от их поверхностной лексической реализации. Это позволяет осуществлять феномен zero-shot и few-shot переноса. При дообучении на ограниченном параллельном корпусе малоресурсного языка модель не выстраивает синтаксические деревья заново, она лишь калибрует проекцию из универсального пространства в конкретную целевую языковую систему (alignment).

Снизить затраты для вычисления при настройке сверхбольших языковых моделей, созданных для исчезающих языков, позволяет более эффективное обучение (PEFT и LoRA), использование этих подходов сокращает количество обучаемых параметров в десятки тысяч раз, сохраняя качество.

Математическая основа метода LoRA [7] базируется на предположении о том, что внутренний ранг матрицы изменения весов в процессе тонкой настройки является очень низким. Пусть матрица весов W_0 имеет размерность d x k. Тогда матрицы B и A будут иметь размерности d × r и r × k соответственно, где ранг r < min (d, k). В процессе прямого прохода градиент вычисляется только для небольших матриц B и A, в то время как гигантская матрица W_0 остается замороженной. Это радикально снижает потребление видеопамяти (VRAM), позволяя обучать модели с миллиардами параметров на одиночных графических ускорителях потребительского класса. Более того, низкоранговая структура выступает в роли сильного регуляризатора: ограничивая емкость обновляемых параметров, LoRA препятствует катастрофическому забыванию (catastrophic forgetting) знаний, полученных на этапе предобучения, и существенно снижает риск переобучения на малом объеме параллельных данных целевого языка. На этапе инференса (inference) матрицы B и A могут быть перемножены и прибавлены к W_0, что исключает любые задержки при вычислении (zero latency overhead).

Что касается объема малоресурсных данных, решением могут служить методы аугментации данных, одним из которых является синтетическая генерация и итеративный обратный перевод (Back-Translation). Этот метод позволяет превратить имеющиеся моноязычные данные в псевдопараллельный корпус [3, с. 490], при использовании данного подхода происходит обучение промежуточной модели перевода, осуществляющей перевод с целевого языка на исходный, затем перевод моноязычного корпуса целевого языка и обучение модели перевода на объединенном корпусе. Трудности сопряжены с трудностью обучения на первом этапе метода, из-за слишком шумного корпуса модель склонна к тому, чтобы допускать эффект распространения ошибок (error propagation); дабы снизить эффект применяется итеративный обратный перевод, в котором качество растет с каждым циклом. Дополнительно используется фильтрация синтетических данных с помощью кросс-языковых метрик вроде LaBSE, отбраковывающая некачественные переводы и обеспечивающая чистоту расширенного корпуса.

Обобщенный анализ эффективности подходов представлен в *табл.* на основе эмпирических данных [6, с. 48].

Таблица

Сравнительный анализ эффективности методов оптимизации Seq2Seq

Метод / Стратегия	Необходимый объем (пар)	Вычислит. затраты	Прирост BLEU
Базовый Transformer	> 1 000 000	Высокие	—
Субсимвольная токенизация (BPE)	от 50 000	Низкие	+2 … +5
Обратный перевод (Back-Translation)	от 10 000	Средние	+4 … +8
Трансферное обучение (mBART)	от 1 000	Очень высокие	+8 … +15
Трансферное обучение + LoRA	от 1 000	Низкие	+7 … +14

Функционирование трансформеров для обучения системы машинного перевода с малоресурсного языка на русский язык можно протестировать с помощью следующих инструментов:

PyTorch: Популярная библиотека для работы с машинным обучением, обеспечивающая широкий набор инструментов для создания и обучения нейронных сетей;

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 позволяет оптимизировать процесс разработки и обеспечить воспроизводимость экспериментов;

система контроля версий: исходный код проекта, включая скрипты для обучения, предобработки данных и вспомогательные утилиты, можно хранить и управлять с использованием GitLab;

язык программирования: проект реализуем на Python версии 3.11. Для управления зависимостями и настройкой среды подходит менеджер пакетов Poetry.

Для обучения и тестирования нейронной сети достаточно обучить модель на выделенном вычислительном кластере, оснащенном 6 GPU NVIDIA A100 с объемом видеопамяти 80 Гб каждая.

В качестве предобученных моделей подойдут:

MADLAD-400-3b;

MADLAD-400-7b;

MADLAD-400-10b;

NLLB-200-3.3b.

Что касается результатов, например, в среднем на одно мансийское слово приходится более двух токенов. Количество предложений, содержащих неизвестные токены, составляет 22 424, при общем датасете из 53 тысяч предложений, что значительно больше, чем у модели MADLAD400, у которой таких предложений всего 32. Основной причиной этого является недостаточный размер словаря модели.

NLLB-200 не охватывает все особенности мансийского языка, в частности, в словаре отсутствует символ «ӈ». Токенизация текстов для этой модели показала, что значительное количество предложений содержит токен UNK. Это указывает на ограничения модели в обработке мансийского языка. Для устранения этой проблемы можно расширить словарь модели NLLB-200, добавив в него всю необходимую символику и токены мансийского языка.

Критическое переобучение, возникающее в классических глубоких Seq2Seq-архитектурах (включая многослойные трансформеры с end-to-end обучением) при дефиците параллельных данных, делает невозможным извлечение универсальных лингвистических репрезентаций и провоцирует OOV-проблемы. Следовательно, для языков с ограниченной цифровой представленностью при проектировании отказоустойчивых систем машинного перевода следует применять каскадную стратегию: предобученные мультиязычные фундаментальные модели, BPE-токенизацию, аугментацию через Back-Translation и адаптацию весов LoRA. Это обеспечивает производственное качество при минимальных ресурсах.

Литература

1. Vaswani A., Shazeer N., Parmar N., et al. Attention is all you need // Advances in Neural Information Processing Systems. 2017. Vol. 30. P. 5998–6008.
2. Lample G., Conneau A. Cross-lingual language model pretraining // Advances in Neural Information Processing Systems. 2019. Vol. 32.
3. Edunov S., Ott M., Auli M., Grangier D. Understanding Back-Translation at Scale. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. ACL, 2018. P. 489–500.
4. Hu E. J., Shen Y., Wallis P., et al. LoRA: Low-Rank Adaptation of Large Language Models. In: International Conference on Learning Representations. Proceedings.com, 2022.

5. Sennrich R., Haddow B., Birch A. Neural Machine Translation of Rare Words with Subword Units. In.: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. ACL, 2016. P. 1715–1725.
6. Бадмаева Е. С., Очиров А. В. Методы машинного перевода для языков с ограниченными ресурсами // Вестник компьютерных и информационных технологий. 2024. №3. С. 45–52.
7. Liu Y., Gu J., Goyal N., et al. Multilingual Denoising Pre-training for Neural Machine Translation // Transactions of the Association for Computational Linguistics. 2020. Vol. 8. P. 726–742.

Seq2Seq architecture for low-resource languages: methods for overcoming data scarcity

R. G. Miftakhova*, N. M. Mylnikov

Ufa University of Science and Technology
32 Zaki Validi st., 450076 Ufa, Republic of Bashkortostan, Russia.

**Email: miftahovar@yandex.ru*

The presented study contains approaches to modifying Seq2Seq models for low-resource language processing scenarios. The key methods of overcoming the “curse of dimensionality” and sparsity of data are considered: algorithms of sub-symbolic tokenization (BPE, SentencePiece); methods of multilingual cross-language transfer (multilingual transfer learning) using high-resource donor languages, as well as methods of synthetic data augmentation (back-translation). Special attention is paid to modern methods of parametrically effective retraining (PEFT). The paper presents a comparative analysis of the productivity of the described approaches, structured according to the size of the available parallel data.

Keywords: machine translation, natural language processing, Sequence-to-Sequence, low-resource languages, transfer learning, reverse translation, LoRA.