Методы:

Название	Ссылка	Год	Комментарий	Этап (просмотрено/исследуется/не притрагивалась)
LLMs for Extremely Low-Resource Finno-Ugric Languages	https://openreview.net/attachment?id=KY3roODQ47&name=pdf	2024	Статья полезна тем, что здесь описаны ресурсы, используемые для обучения (базис языков для pre-training: **русский, английский и латышский** (доля в данных по 12% каждый) и **финский с эстонским** по 32%). Если не будем предобучать из-за нехватки ресурсов, то стоит искать модели с таким бэком. Вряд ли стоит рассматривать ллмки, у нас нет таких ресурсов, но вывод по bleu такой: ru->komi ~ 14.5	Просмотрено
Machine Translation for Low-resource Finno-Ugric Languages	https://aclanthology.org/2023.nodalida-1.77.pdf	2023	Также поддерживает идею перевода монокорпуса и предлагает модели для “хорошей стартовой точки” для файнтюна (nllb и m2m). Использованы: M2M-100, 1.2 billion parameters (multi-lingual neural machine translation model); Для обучения также использовался Fairseq framework. Пример их finetune: https://github.com/TartuNLP/m2m-100-finetune	Просмотрено

			Стоит также отметить, что они увеличивали вокабуляр и размер матрицы эмбеддингов с помощью этих скриптов. Аккуратнее с библейскими данными, люди пишут, что переобучилось на них.	
NEURAL MACHINE TRANSLATION FOR LOW RESOURCE LANGUAGES	https://arxiv.org/pdf/2304.07869	2023	Используют перевод по словам для ускорения претрейнинга биязычной модели. Для претрейна биязычной модели используется masked language model (MLM) на моноязычных данных на обоих языках. Предлагают добавлять третий “язык” с переводом по словам, чтобы модель училась сопоставлять слова из обоих языков: “с целью MLM предсказать замаскированное английское слово, модель может учитывать как английские, так и иностранные слова в предложении “третьего языка”, и наоборот”. Идею для получения параллельных предложений с помощью меры жаккарда стоит попробовать на данных из Луима Серипос. Тоже использовался фреймворк Fairseq. Использовали самописный focal loss, описанный в https://aclanthology.org/2020.findings-emnlp.276.pdf (я также нашла реализацию https://github.com/vyraun/long-tailed/blob/main/fairseq/criterions/focal_loss.py) Тренировали mbart-cc25.	Просмотрено

Machine Translation for Livonian: Catering to 20 Speakers	https://aclanthology.org/2022.acl-short.55.pdf	2022	Подводка к проекту OPUS с opensource кодом. Специализируются на лоу ресурс языки. Предлагаемая архитектура: 6 слоев энкодеров и декодеров, 8 attention heads на каждый слой, word embeddings и hidden layers размера 512, dropout на 0.3, максимальная длина предложения - 128 символов. Обучение проходило с помощью FairSeq тулы (pytorch) - https://github.com/facebookresearch/fairseq Их готовая модель для ливонского (как пример): https://huggingface.co/tartuNLP/liv4ever-mt Генерация доп данных делается с помощью УЖЕ обученной модели, просто генеря лучшей моделью перевод монокорпуса.	Просмотрено
Low-Resource Machine Translation Training Curriculum Fit for Low-Resource Languages	https://arxiv.org/pdf/2103.13272	2021	Использовалась модель https://github.com/facebookresearch/XLM, потом был пре-трейн двуязычной LM на задаче MLM на монокорпусе. Также предлагается ввести третий язык, чтобы выровнять эмбеддинги английского и иностранного. Дальнейшая стадия включает в себя unsupervised пре-трейн NMT (энкодер и декодер - предобученная до этого модель) на бек-транслейшн монокорпуса. По доп данным также смотрят по Жаккарду. Также для оценки сопоставимых данных	Просмотрено

			используется Ratio Margin-based Similarity Score. Авторы отмечают, что используют 1гпу на 32Гб.	
Understanding Back-Translation at Scale	https://aclanthology.org/D18-1045.pdf	2018	Создание синтетических данных для лоу ресурс	Не просмотрено
Exploring Diversity in Back Translation for Low-Resource Machine Translation	https://aclanthology.org/2022.deeplo-1.8.pdf	2022	Тоже статья о работе с данными	Не просмотрено
Improving Low-Resource Neural Machine Translation with Filtered Pseudo-parallel Corpus	https://aclanthology.org/W17-5704.pdf	2017	Берется предложение, выполняется перевод из таргета (моноязычные данные) в сурс-язык. получаем сурс_синтетику. переводим обратно из сурса_синт в таргет_синт. сравниваем таргет и таргет_синт по схожести. сортируем данные по метрике схожести, выбираем некоторое пороговое значение, все что ниже его - отсекаем. получаем синетические данные для обучения. + используют бут	Просмотрено

Предлагаемые модели для безлайна:

Название	Ссылка	Комментарий	
smugri3-finno-ugric-nmt	https://huggingface.co/tartuNLP/smugri3-finno-ugric-nmt	Весит 15Гб. Есть Манси, Ханты	
smugri3_14-finno-ugric-nmt	https://huggingface.co/tartuNLP/smugri3_14-finno-ugric-nmt	Весит 9Гб. Есть Манси, Ханты	
tartuNLP/m2m-1_2B-finetune-finno-ugric-bt2	https://huggingface.co/tartuNLP/m2m-1_2B-finetune-finno-ugric-bt2	Весит 5Гб. Файнтюн на финно-угрик языки.	
m2m-1_2B-finetune-finno-ugric-bt2	https://huggingface.co/tartuNLP/m2m-1_2B-finetune-finno-ugric-bt2	Нет описания, авторы одни и те же с моделями выше, но весит чуть меньше 5Гб	
Opus-MT	https://github.com/Helsinki-NLP/Opus-MT-train?tab=readme-ov-file		
facebook/nllb-200-distilled-1.3B	https://huggingface.co/facebook/nllb-200-distilled-1.3B/tree/main	Весит 5.5Гб. Есть примеры файнтюна в переводчик.	
facebook/m2m100_418M	https://huggingface.co/facebook/m2m100_418M/tree/main	Весит 2Гб. Вроде эту модель использовали авторы кейса. Уже обучена как переводчик (на многие языки).	Большая вариативность глубины модели.
michaelfeil/ct2fast-m2m100_1.2B	https://huggingface.co/michaelfeil/ct2fast-m2m100_1.2B	Несмотря на большое кол-во параметров, модель квантизирована и имеет сравнительно небольшой вес.	

tartuNLP/m2m100_418M_smugri	https://huggingface.co/tartuNLP/m2m100_418M_smugri/tree/main	Весит 2Гб. Предобучена на Финно-угорских языках.	
facebook/mbart-large-50	https://huggingface.co/facebook/mbart-large-50	Весит 2.5Гб. Много решений с использованием mBART.	Это предварительно обученная модель, которая в первую очередь предназначена для точной настройки в задачах перевода.
mBART	https://huggingface.co/facebook/mbart-large-cc25	Весит 2.5Гб. Много решений с использованием mBART.	Пример тюнинга mBART

Полезные гайды:

Название	Ссылки	Комментарий
Перевод	https://huggingface.co/docs/transformers/v4.17.0/en/tasks/translation https://medium.com/@tskumar1320/how-to-fine-tune-pre-trained-language-translation-model-3e8a6aace9f	Пример обучения из коробки, можно попробовать для быстрого безлайна