В данном файле представлен research part нашего проекта.

Методы:

Название	Ссылка	Год	Комментарий	Этап
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);	Просмотрено

			Для обучения также использовался Fairseq framework. Пример их finetune: https://github.com/TartuNLP/m2m-100-finetune 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	Просмотрено

			.pdf (я также нашла реализацию https://github.com/vyraun/long-tailed/blob/main/fairseq/criterions/focal_loss.py) Тренировали mbart-cc25.	
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	Просмотрено

			декодер - предобученная до этого модель) на бек-транслейшн монокорпуса. По доп данным также смотрят по Жаккарду. Также для оценки сопоставимых данных 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.	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 бейзлайна:

Название	Ссылка	Комментарий	
smugri3-finno-ugric-nmt	https://huggingface.co/tartuNLP/smugri3-finno-ugric-nmt	Весит 15Гб. Есть Манси, Ханты	Не смогли использовать из-за требовательности к ресурсам. Метрика BLEU для наших данных равна 0.0039
smugri3_14-finno-ugric-nmt	https://huggingface.co/tartuNLP/smugri3_14-finno-ugric-nmt	Весит 9Гб. Есть Манси, Ханты	Не смогли использовать из-за требовательности к ресурсам.
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.	Не смогли использовать из-за требовательности к ресурсам.
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	Не смогли использовать из-за требовательности к ресурсам.
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.	Смогли запустить и затюнить параметры, результаты оказались хуже модели m2m100.
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. Смогли запустить, стала основной архитектурой.

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.	Смогли запустить с помощью LoRA, однако такое обучение не оказалось успешным - лосс меньше 6 не падал.
tartuNLP/m2m100_418M_smugri	https://huggingface.co/tartuNLP/m2m100_418M_smugri/tree/main	Весит 2Гб. Предобучена на Финно-угорских языках.	Смогли запустить, но результаты оказались похуже m2m100.
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. результаты оказались такими же, как и у m2m100, однако училась дольше.
mBART	https://huggingface.co/facebook/mbart-large-cc25	Весит 2.5Гб. Много решений с использованием mBART.	Пример тюнинга mBART

Адаптеры - Полезные гайды:

Название	Ссылка	Коммент
Efficiently train Large Language Models with LoRA and Hugging Face	https://github.com/roy-sub/LLM-FineTuning/blob/main/1.Efficiently_train_Large_Language_Models_with_LoRA_and_Hugging_Face.ipynb	Пример использования LoRA на seq2seq задаче
Кто же такая это ваша LoRA	https://habr.com/ru/articles/747534/	Небольшая статейка на хабре для тех, кто не знал о Лоре

Результаты запуска экспериментов:

Имя	Модель	Размер модели	Тип модели	Ссылка	Железо для обучения	Время на обучение	На чем обучен	BLEU	chrf	Комментарий	Ссылка на чекпоинт
Никита	facebook/mbart-large-50-many-to-many-mmt	610m	моно	https://huggingface.co/facebook/mbart-large-50-many-to-many-mmt	P100	3 эпохи (12171 шагов), 5 часов 30 минуты	на параллельном корпусе	21.7 (rus->mansi)	52		
Анна	facebook/m2m100_418M	418m	моно	https://huggingface.co/facebook/m2m100_418M/tree/main	P100	3 эпохи (12171 шагов), 4 часа 30 минут	на параллельном корпусе	21.7 (rus->mansi)	52	стоит дать больше времени сойтись	https://disk.yandex.ru/d/5B-fhMTjiP1ABA
Лера	facebook/mbart-large-cc25	2.5гб	моно	https://huggingface.co/facebook/mbart-large-cc25	P100	3 эпохи (12171 шагов), 5 часов 33 минуты	на параллельном корпусе			метрики упали	
Никита	m2m-bilingual	418m	моно	https://www.kaggle.com/models/nsgorbunov/2ndcheckpoint/PyTorch/default/1	P100	3 эпохи	на параллельном корпусе рус+манси	22.6 (rus->mansi & mansi->rus)	53.8		
Анна	tartuNLP/m2m100_418M_smugri	418m	моно	https://huggingface.co/tartuNLP/m2m100_418M_smugri/tree/main	P100	3 эпохи (12171 шагов), 4 часа 20 минут	на параллельном корпусе	19.12 (rus->mansi)	50		не сохранился
Анна	facebook/m2m100_418M	418m	моно	https://huggingface.co/facebook/m2m100_418M/tree/main	P100	2 эпохи (19к шагов), 7 часов 30 минут	на монокорпусе			loss: 1.1	https://disk.yandex.ru/d/zXFYJf9CjzYG4Q
Анна	facebook/m2m100_418M + mlm + finetune	418m	моно	https://disk.yandex.ru/d/zXFYJf9CjzYG4Q	P100	3 эпохи (12171 шагов), 4 часа 20 минут	на параллельном корпусе	21.2 (rus->mansi)	51.2		https://disk.yandex.ru/d/MIzo6jE0gDRIjQ

Анна	facebook/m2m100_418M	418m	билингв	https://huggingface.co/facebook/m2m100_418M/tree/main	P100	1 эпоха (8114 шагов), 3 часа	на параллельном корпусе рус+манси	17.6 (rus->mansi & mansi->rus)	45.3	Нужно больше эпох	https://disk.yandex.ru/d/Q8yzRnaPKT1SlA
Лера	facebook/m2m100_418M	418m	билингв	https://disk.yandex.ru/d/Q8yzRnaPKT1SlA	P100	2ая эпоха аниной модели, 3 часа	на параллельном корпусе рус+манси	rus -> mansi 20.88, mansi -> rus 21.1	rus -> mansi 50.7, mansi -> rus 46.86		https://www.kaggle.com/models/nsgorbunov/2ndcheckpoint/PyTorch/default/1
Лера	facebook/m2m100_418M	418m	билингв	https://www.kaggle.com/models/nsgorbunov/2ndcheckpoint/PyTorch/default/1	P100	3 эпоха аниной модели	на параллельном корпусе рус+манси	rus -> mansi 22.6, mansi -> rus 22.84	rus -> mansi 53.88, mansi -> rus 49.07		https://www.kaggle.com/datasets/riapush/3checkpoint
Лера	facebook/m2m100_418M	418m	билингв	https://www.kaggle.com/datasets/riapush/3checkpoint	P100	1 эпоха	на параллеьном корпусе синтетики	rus -> mansi 21.38, mansi -> rus 21.4	rus -> mansi 53.08, mansi -> rus 48.02	метрики упали :(	
Лера	facebook/m2m100_418M	418m	билингв	https://www.kaggle.com/datasets/riapush/3checkpoint	P100	2 эпоха	на библии и словарях	rus -> mansi , mansi -> rus	rus -> mansi , mansi -> rus		