Ни одного языка в стороне:

Масштабирование центрированного на человеке машинного перевода

Команда NLLB:

Марта Р. Коста-Жуссà, Джеймс Кросс, Онур Челеби, Маха Эльбаяд, Кеннет Хеффилд, Кевин Хеффернан, Элахе Калбасси, Джэнис Лэм, Дэниэл Лихт, Жан Майяр, Анна Сан, Скайлер Ванг, Гийом Вензек, Аль Янгблад, Бапи Акула, Лоик Барро, Габриэль Мехия Гонсалес, Прангтип Хансанти, Джон Хоффман, Семарли Джарретт, Каушик Рам Садагопан, Дирк Роу, Шеннон Спруит, Чау Тран, Пьер Андреуси, Неджип Фазиль Айаны, Шрути Бхосале, Сергей Эдунов, Анджела Фаны, Синтия Гао, Веданудж Госвами, Франциско Гузман, Филипп Кёни, Александр Мурачко, Кристоф Роперс, Сафия Салими, Хольгер Швенк, Джефф Ванг

Meta AI, Университет Калифорнии в Беркли, Университет Джона Хопкинса

Аннотация:

Стремясь к искоренению языковых барьеров в глобальном масштабе, машинный перевод стал ключевым направлением исследований в области искусственного интеллекта. Однако эти усилия сосредоточены на небольшом подмножестве языков, оставляя без внимания подавляющее большинство языков с ограниченными ресурсами. Что нужно для преодоления барьера в 200 языков, обеспечивая при этом безопасные и качественные результаты, и не забывая об этических соображениях? В рамках проекта "No Language Left Behind" мы взялись за эту задачу, сначала осмыслив необходимость поддержки перевода для языков с ограниченными ресурсами через исследовательские интервью с носителями языка. Затем мы создали наборы данных и модели, нацеленные на сокращение разрыва в производительности между языками с низкими и высокими ресурсами. В частности, мы разработали модель условного вычисления, основанную на разреженной смеси экспертов (Sparsely Gated Mixture of Experts), обученную на данных, полученных с помощью новых и эффективных методов добычи данных, адаптированных для языков с ограниченными ресурсами. Мы предлагаем несколько архитектурных и учебных улучшений для противодействия переобучению при обучении на тысячах задач. Критически мы оценили производительность более чем 40,000 различных направлений перевода, используя эталонный тест, переведенный людьми, Flores-200, и совместили человеческую оценку с новым эталоном токсичности, охватывающим все языки в Flores-200, для оценки безопасности перевода. Наша модель достигает улучшения на 44% по метрике BLEU относительно предыдущего передового уровня, закладывая важные основы для создания универсальной системы перевода. Наконец, мы открываем доступ ко всем описанным в этой работе материалам, которые доступны по адресу: https://github.com/facebookresearch/fairseq/tree/nllb.

Содержание

1 Введение
2 Человеко-центрированный перевод языков с ограниченными ресурсами
2.1 Исследование дизайна исследования на основе эксплораторских интервью
2.2 No Language Left Behind: Руководящие принципы
3 Языки
4 Создание профессионально переведенных наборов данных: FLORES-200 и NLLBSeed
4.1 FLORES-200
4.2 Набор данных NLLB Seed
4.3 Многодоменный набор данных NLLB
4.4 Заключение 5 Автоматическое создание тренировочных данных для перевода на сотни языков
5.1 Идентификация языка
5.2 Сбор и очистка монолингвистических данных масштаба
5.3 Добыча битекстов для языков с ограниченными ресурсами
5.4 Заключение
6 Моделирование
6.1 Предварительные работы
6.2 Условные вычисления для массового мультиязычного машинного перевода
6.3 Стратегии самонадзора на крупномасштабных монолингвистических корпусах
6.4 Увеличение данных
6.5 Инициализация моделей с использованием NLLB-Seed
6.6 Заключение
7 Оценка
7.1 Автоматическая оценка
7.2 Человеческая оценка
7.3 Токсичность
7.4 Заключение
8 Всё вместе
8.1 Подготовка данных
8.2 Подготовка модели
8.3 Результаты на Flores-200
8.4 Обобщение вне области: Производительность в доменах, не относящихся к Flores-200
8.5 Анализ NLLB-200
8.6 Сделать большие модели доступнее через дистилляцию
8.7 Эффективное включение языков с несколькими алфавитами и связанными лангойдами
8.8 Экологическое влияние NLLB 9 No Language Left Behind: Социальное влияние и заключительные мысли
9.1 Расширение доступа к информации
9.2 Двуликость цифрового участия

9.3 Будущее NLLB: Коллективная ответственность
10 Вклад
11 Благодарности
A Языки
B Оценка
C Данные
D Моделирование
E Всё вместе
F Карточка модели - NLLB-200
G Карточка данных для данных NLLB-Seed
H Карточка данных для многодоменных данных NLLB
I Карточка данных для метаданных добытых битекстов

Введение

В научно-фантастическом романе Джека Вэнса (1977) "Глаза загробного мира" главный герой Кугел сталкивается с волшебником, который поручает ему задачу. Чтобы помочь ему, волшебник дарит Кугелу магическое устройство: "Чтобы облегчить твою речь, я наделяю тебя этим инструментом, который связывает все возможные слова с любой мыслимой системой значений."

Перенесемся на полвека вперед: теперь мы знаем, что магическое устройство Кугела на самом деле является машинным переводом. Машинный перевод, задуманный как вычислительные системы для перевода текстов с одного языка на другой, существует с 1940-х годов, но его недавний переход от статистических (Brown и др., 1993; Koehn, 2009; Lopez, 2008) к нейронным системам вывел технологию на новые рубежи (Bahdanau и др., 2015; Cho и др., 2014; Kalchbrenner и Blunsom, 2013; Wu и др., 2016). Этот сдвиг не только стремительно улучшил качество перевода, но и расширил применение машинного перевода в новых областях. Сегодня машинный перевод влияет на то, как люди по всему миру общаются, работают, путешествуют, учатся, получают доступ к информации и многое другое (Khoong и Rodriguez, 2022; Koehn и Germann, 2014; Lee, 2020).

Хотя машинный перевод продолжает развиваться, его плоды распределены неравномерно (Fan и др., 2020). На самом деле, подавляющее большинство улучшений, достигнутых в машинном переводе за последние десятилетия, были сделаны для языков с большими ресурсами, или языков, для которых доступно большое количество учебных данных в цифровом виде. Например, те, кто общается на английском, французском, немецком или русском — языках, которые давно получают институциональные инвестиции и имеют доступные данные — получают значительно больше пользы от развития машинного перевода, чем те, кто говорит на каталанском, ассамском, лигурийском или киньяруанда. Многие из этих языков привлекают меньше внимания и ресурсов, несмотря на то, что большинство языков, на которых говорят в мире сегодня, являются языками с ограниченными ресурсами (Joshi и др., 2020). Многие из этих языков остаются вне поля зрения исследователей по совокупности причин, включая ограничения, вызванные прошлым финансированием (или его отсутствием), исследовательскими нормами, организационными приоритетами и западноцентризмом, если назвать лишь некоторые из них. Без усилий по исправлению курса, значительная часть интернета может остаться недоступной для носителей этих языков. Исследования показывают, что, хотя только 25,9% пользователей интернета говорят на английском, 63,7% всех веб-сайтов написаны на английском (следующим в списке идет русский с 6,8%; Richter, 2022). Для многих сообществ, говорящих на языках с ограниченными ресурсами, "Многоязычный Интернет" (Zuckerman, 2008), важное средство, которое могло бы способствовать доступу к образованию и социальной мобильности, остается недостижимым, поскольку веб-долгое время отдавал приоритет контенту, ориентированному на носителей языков с большими ресурсами.

Расширение машинного перевода на большее количество языков с ограниченными ресурсами также сдерживается техническими проблемами (Haddow и др., 2022). По сравнению с их коллегами с большими ресурсами, данные для обучения языков с ограниченными ресурсами дороги и логистически сложно достать (Kuwanto и др., 2021; Nekoto и др., 2020; Orife и др., 2020). Без достаточного количества учебных данных стандартные методы могут не выдержать проверку временем при возникновении новых требований. Эти препятствия стали еще более заметными, поскольку популярность методов, требующих большого количества данных, таких как масштабное предварительное обучение и масштабирование моделей, стала мейнстримом (Conneau и Lample, 2019; Conneau и др., 2020; Kenton и Toutanova, 2019; Radford и др., 2019).

Для преодоления этих барьеров значительная часть существующих работ по переводу для языков с ограниченными ресурсами сосредоточена на использовании многоязычных систем, или моделей, способных работать с несколькими языками. Эти модели имеют преимущество межъязыкового переноса (Nguyen и Chiang, 2017; Zoph и др., 2016),

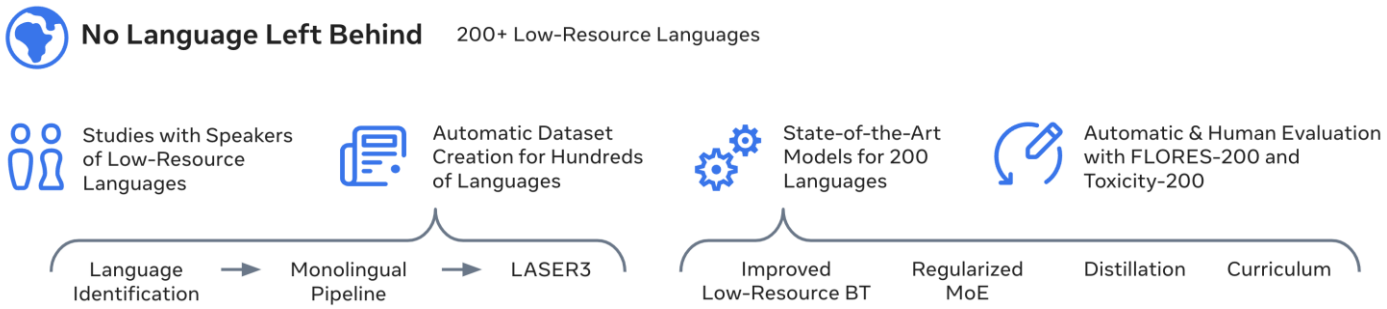


Рисунок 1: Инициатива No Language Left Behind: Наше усилие по переводу с ограниченными ресурсами сосредоточено на четырех основных принципах. (1) Мы стремимся понять проблему перевода с ограниченными ресурсами с точки зрения носителей языка. (2) Мы изучаем способы автоматического создания тренировочных данных для перевода языков с ограниченными ресурсами к уровню высоких ресурсов. (3) Мы используем эти данные для создания передовых моделей перевода. (4) Мы оцениваем каждый язык, который намечаем переводить.

разрешение родственным языкам учиться друг у друга (Arivazhagan и др., 2019; Fan и др., 2020; Zhang и др., 2020). В то время как мультиязычные модели продемонстрировали многообещающее улучшение производительности по сравнению с двуязычными моделями (Tran и др., 2021), обеспечение представления сотен языков при сохранении высокого качества перевода остается открытой областью исследований. Другая стратегия, направленная на смягчение проблемы недостатка ресурсов, заключается в сборе большего объема языковых данных. Некоторые из этих попыток сосредоточены на сборе человеческих переводов, в то время как другие используют крупномасштабный майнинг данных и монолингвистические трубопроводы для консолидации данных, найденных в сети (Bañón и др., 2020; Karakanta и др., 2018; Ramesh и др., 2022; Schwenk и др., 2021b). Последние техники часто страдают от шума и предвзятости, что затрудняет проверку качества созданных наборов данных (Kreutzer и др., 2022). Наконец, разработка моделей перевода для языков с ограниченными ресурсами требует наличия высококачественных, человекопереведенных оценочных бенчмарков. Наборы данных, такие как Flores-101 (Goyal и др., 2022), работают в этом направлении, но их покрытие ограничено 100 языками. В этой статье мы задаемся вопросом: что нужно сделать, чтобы удвоить охват языков большинства существующих моделей перевода, обеспечивая при этом высокое качество и безопасные переводы? Конкретнее, как использовать человекоцентрический подход (Robertson и др., 2021), чтобы создать грамотные, сохраняющие значение переводы для более чем 200 языков, многие из которых относятся к классу языков с ограниченными ресурсами, которые остаются недооцененными существующими технологиями перевода? И как можно это сделать, минимизируя потенциальный вред от катастрофических и токсичных переводов, создаваемых нейронными моделями машинного перевода - редкими случаями, которые, тем не менее, имеют значительное негативное воздействие на пользователей? Мы берем на себя этот вызов в рамках инициативы No Language Left Behind (NLLB). Мы начинаем с создания Flores-200, многоязычного набора данных с множеством-к-множеству, что позволяет нам измерять качество перевода в любом из 40,602 общих направлений перевода. Мы разработали технику кодирования предложений на основе дистилляции, LASER3 (Heffernan и др., 2022), которая помогла нам майнить веб-данные для создания параллельных наборов данных для языков с ограниченными ресурсами. Используя как майнингованные данные, так и набор данных с человеческими переводами, мы обучили мультиязычные модели экспертов-смесей с передовым уровнем производительности. Несмотря на удвоение количества языков, наша конечная модель демонстрирует улучшение на 40% по сравнению с предыдущими достижениями на Flores-101. Для выявления и предотвращения потенциально вредных переводов, создаваемых моделями перевода, мы создали набор данных токсичных слов для всех 200 языков путем комбинирования автоматической и человеческой оценок.

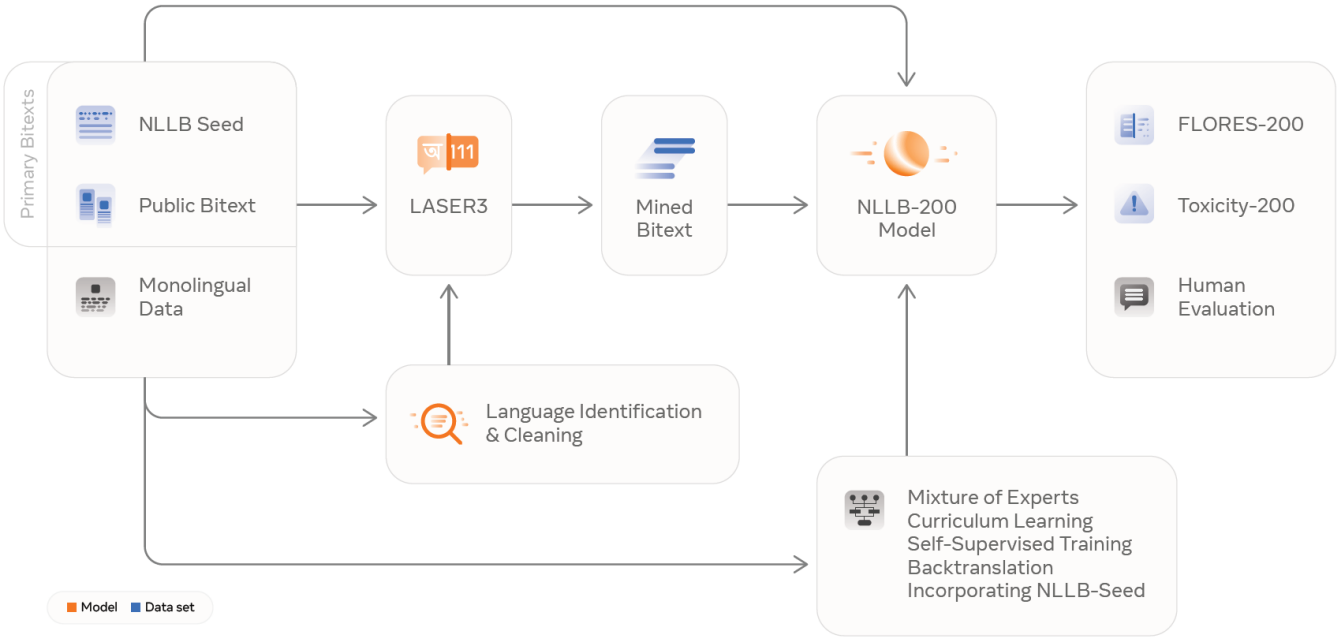


Рисунок 2: **Как компоненты взаимодействуют, общий обзор**: Мы изображаем технические компоненты инициативы No Language Left Behind и их взаимосвязь. Мы показываем взаимодействие данных, как данные используются в разрабатываемых нами моделях (оранжевый цвет), и как модели оцениваются. Датасеты, изображенные синим цветом, являются новыми наборами данных, созданными в рамках инициативы No Language Left Behind.

Мы предложили и провели человеческие оценки на множестве языков, которые охватывают наши модели, помимо стандартных автоматических метрик, чтобы получить качественное представление о влиянии перевода. Кроме того, помимо создания этих моделей, мы также рефлектируем на процесс их создания, анализируя риски и выгоды нашего исследования с общественной точки зрения. Мы открываем исходный код всех бенчмарков, данных, скриптов и моделей, описанных в наших усилиях, чтобы поддержать дальнейшие исследования. Кроме того, мы сосредотачиваемся на практической применимости нашей работы для сообществ с ограниченными ресурсами. Мы применяем наши техники для предоставления поддержки в переводе редакторам Википедии, что позволяет им более эффективно создавать новые статьи на языках, которые не поддерживаются другими системами перевода.

Остальная часть статьи структурирована следующим образом, с обзором на Рисунок 2: Раздел 2 описывает открытые проблемы в переводе с ограниченными ресурсами и анализирует широкое использование систем перевода. Раздел 3 представляет языки, на которые мы сосредоточились, и объясняет, как мы пришли к этому набору языков. Раздел 4 подводит итоги процесса создания наборов данных Flores-200 и NLLB-Seed + NLLB-MD для начального перевода с анализом качества. Раздел 5 обзорно рассматривает создание монолингвистических и майнинговых двуязычных данных, что позволяет создавать модели для сотен языков. Раздел 6 подробно описывает различные техники моделирования, разработанные для улучшения производительности языков с ограниченными

ресурсами. Раздел 7 отслеживает автоматическую и человеческую оценку наших переводов, включая обнаружение катастрофических и токсичных переводов. Мы интегрируем упомянутые наборы данных и техники в модель NLLB-200, которая в настоящее время поддерживает 202 языка, и анализируем ее качество и производительность в Разделе 8. Мы заключаем в Разделе 9, где мы рефлектируем о социальном влиянии наших исследований и рассматриваем будущие возможности и вызовы. Надеемся, что наш вклад поможет будущим исследователям, которые, подобно нам, стремятся увидеть волшебное устройство Кугеля - машинный перевод, охватывающий все языки, - стать реальностью, а не только концептуальной химерой.

Для доступности нашей работы сообществу мы предоставляем следующее в открытый доступ:

• Наборы данных, переведенные людьми Flores-200: Оценочный набор данных на 204 языках NLLB-Seed: Начальные тренировочные данные на 39 языках NLLB-MD: Начальные данные в различных областях на 6 языках для оценки обобщения Toxicity-200: Словники для обнаружения токсичности на 200 языках

• Инструменты для создания масштабных битекстовых наборов данных Идентификация языков для более чем 200 языков LASER3: Сентенциальные энкодеры для идентификации выровненных битекстов на 148 языках stopes: Библиотека для добычи данных, которая может использоваться для обработки и очистки монолингвистических данных, а затем создания выровненных битекстов Восстановление тренировочных данных: Скрипты для воссоздания наших тренировочных данных

• Модели перевода, охватывающие 202 языка NLLB-200: Модель смеси экспертов с разреженными воротами на 54.5 миллиарда параметров 3.3 миллиарда и 1.3 миллиарда плотных трансформерных моделей 1.3 миллиарда и 600 миллионов плотных трансформерных моделей, дистиллированных из NLLB-200 Скрипты для тренировки и генерации для воспроизведения наших моделей

2. Перевод с участием человека для языков с ограниченными ресурсами

Для определения нашей цели предоставления высококачественного перевода для сотен языков мы вначале исследуем важность этого исследования для тех, кто наиболее важен для нас: сообществ с ограниченными языковыми ресурсами. Вдохновленные концепцией "Значимо Чувствительного Дизайна" (Friedman и Hendry, 2019; Van Der Hoven и Manders-Huits, 2020), мы выделяем интересы и ценности на уровне сообществ как основополагающие для нашего исследования. Принятие этой рамки побуждает нас начать с людей и приоритизировать их взаимодействие с технологиями, с прямым акцентом на этические и социальные аспекты (Mukhija и др., 2021). Чтобы понять, как носители языков с ограниченными ресурсами воспринимают машинный перевод, мы провели исследование интервью с 44 носителями таких языков. В качестве заинтересованных сторон, которые могут быть затронуты проектом No Language Left Behind (NLLB), их вклад помог нам представить обещания, в которые многие верят, что машинный перевод может принести их сообществам. Они выразили свою осторожную оптимистичность и предложили конкретные идеи о способах максимизации социальных выгод при минимизации рисков. Более того, многие участники интервью нарисовали наглядные картины культурной и политической обстановки, в которой существуют их языки, сплетение языка и социальных опытов, и как NLLB потенциально может изменить культурное статус-кво.

2.1 Основы исследования на основе исследовательских интервью

Мы разработали полуструктурированный протокол интервью с целью изучения потребностей и опасений носителей языков с ограниченными ресурсами по отношению к машинному переводу. Хотя языки с ограниченными ресурсами могут считаться таковыми по различным причинам, включая малое количество исследований, цифровизацию или обучение (Cieri и др., 2016; Magueresse и др., 2020), для целей исследования мы определяем их как языки, у которых на момент исследования было менее 1 миллиона предложений общедоступных примеров переводов. В ходе интервью были зафиксированы широкий спектр отношений и пониманий, включая использование и применение языков с ограниченными ресурсами, воспринимаемую ценность технологии перевода и принципы разработки систем перевода.

Всего наше набор участников составили 44 носителя языков с ограниченными ресурсами с различными фонами, в возрасте от 23 до 58 лет. Они представляли 36 языков, распределение по которым следующее: 5 языков преимущественно говорятся в Северной Америке, 8 в Южной Америке, 4 в Европе, 12 в Африке и 7 в Азии. Хотя наша выборка разнообразна по расе, образованию и местоположению, большинство участников — иммигранты, проживающие в США и Европе, и около трети из них (n = 17) работают в сфере технологий. Все интервью проводились удаленно через видеоконференц-связь. В среднем интервью длились 1.5 часа. Две трети интервью были записаны и транскрибированы. Для не записанных интервью два исследователя вели обширные заметки. После сбора всех 44 интервью ответы систематически кодировались, что позволило выявить основные темы и идеи.

Мы признаем, что отбор носителей языков с ограниченными ресурсами из диаспорных контекстов имеет свои ограничения. Во-первых, как иммигранты, их точки зрения могут не полностью отражать чувства их сообществ на родине. Тем не менее, некоторые исследователи аргументируют, что в технологически неразвитых странах, где находится множество языков с ограниченными ресурсами, люди склонны рассматривать технологии более оптимистично и амбициозно, чем те, кто живет в странах с более высоким уровнем технологического развития (Kapania и др., 2022; Kozyreva и др., 2021; Sambasivan, 2021; Sambasivan и др., 2021). Таким образом, подвергаясь критическим технологическим дискурсам (особенно в последнее время), многие из наших интервьюируемых могли бы получить более сбалансированное видение рисков, связанных с технологическим прогрессом. Более того, исследователи иммиграции напоминают нам, что глобальное движение сегодня представляет собой транснациональный процесс, в котором те, кто находится в принимающих обществах, поддерживают культурные связи с теми, кто остается в странах отправления через различные коммуникативные и медийные платформы (Baldassar и др., 2016; Levitt и Jaworsky, 2007; Levitt и Lamba-Nieves, 2011). Поскольку в наших интервью мы нашли сильные доказательства таких процессов, мы уверены, что наши участники находятся в уникальной позиции для критического и осведомленного обсуждения социологических оснований своих языков.

Излишнее включение работников в сфере технологий может ввести в исследование форму селективного искажения. Более конкретно, исследования показывают, что работники в сфере технологий, обладая статусом "внутреннего" эксперта, вероятно, будут отстаивать техно-оптимизм — положительное отношение к технологическому развитию (McLennan, 2016). В то время как такой эффект нельзя

недооценивать, личное увлечение работников в сфере технологий технологическими практиками на самом деле может обеспечить им критическую рефлексию, которую мы стремились использовать. Как предполагалось, многие участники рассматривали выгоды нашего исследования, но они также были заинтересованы в выделении потенциальных рисков, которые такое вмешательство может повлечь за собой для их собственных языковых сообществ. Эти тонкие перспективы имели важное значение для формирования наших исследовательских процессов и процедур.

2.1.1 Почему мы должны отдавать приоритет языкам с ограниченными ресурсами?

Язык не только способ общения между людьми, но также передает культуру, историю и самоидентификацию (Демикелис и Вейбулл, 2008; Холл, 2013). В качестве связующего элемента язык способствует формированию сообщества, продолжая традиции и наследие общего народа. Несмотря на то что многие из наших интервьюированных носителей языков с ограниченными ресурсами владеют английским языком, почти все они подчеркивают, что их родной язык остается основополагающей частью их личной идентичности. Сравнивая себя и свои сети с домашними условиями, более половины участников нашего исследования жалеют, что без устойчивых усилий по приоритизации использования и применения их родных языков, многие из них столкнутся с угрозой исчезновения в будущем.

Упадок родного языка и культуры. Опасения о том, что языки с ограниченными ресурсами могут находиться в состоянии упадка, прозвучали на протяжении интервью. Такие утверждения обычно относят упадок к двум причинам: культурной и экономической. Культурная теория предполагает, что по мере того как все больше аспектов нашей жизни становятся цифрово-средственными, длительное воздействие контента, находящегося в сети и на платформах социальных медиа (например, YouTube, Facebook), приводит к приоритизации языков с высокими ресурсами. В этом контексте подчеркивается западная эпистемология и идеи передовых пониманий (Нурулла, 2008). Как отметили некоторые интервьюируемые, культурное доминирование Запада оказывает сильное давление на местные медиа-продукции. По мере того как носители языков с ограниченными ресурсами устремляются к книгам, фильмам и контенту социальных медиа, адаптированным для аудиторий с высокими ресурсами, интерес к продуктам культуры на языках с ограниченными ресурсами может быть вытеснен. Без постоянной аудитории культурные продукты на языках с ограниченными ресурсами рискуют утратить свое место.

Другая группа атрибутирует упадок языков с ограниченными ресурсами влиянию глобальной политической экономики. Для многих носителей языков с ограниченными ресурсами, приходящих из развивающихся стран, язык с высокими ресурсами, такой как английский, рассматривается как средство для глобальной конкурентоспособности и социальной мобильности. Приоритизация лингва франка глобальной экономики означает направление больших ресурсов на образование на английском и привязку местных сообществ к потребностям знаниевой экономики, преимущественно управляемой западными потребностями. Рассматривая это через призму нулевой суммы, многие интервьюируемые считают, что продвижение английского языка может означать дальнейшее периферийное положение родных языков в общественной жизни. Под давлением таких факторов статус многих языков с ограниченными ресурсами находится под угрозой дальнейшего отступления.

Замечая эти тенденции, многие носители языков с ограниченными ресурсами напоминают нам, что машинный перевод может стать критическим инструментом в продвижении сохранения языка и культуры. Как выразился носитель языка игбо, улучшение машинного перевода для его родного языка позволит большему числу людей создавать культурные знания на этом языке, добавив позднее, что веб-сайты, такие как Wikipedia, могут стать важной платформой, позволяющей другим узнавать о истории и практиках его культуры. Отзвучав такие аргументы, другой участник интервью подчеркивает важность двунаправленного обучения, отмечая, что возможность перевода позволяет людям, не говорящим на их языке, читать и понимать статьи Wikipedia о их культуре, что дополнительно мотивирует других писателей продолжать творить. Таким образом, двунаправленное обучение не только освещает сложные взаимоотношения между машинным переводом и сохранением культуры, оно предоставляет возможность нарушить укрепляющуюся природу западно-центричного распространения знаний. Центральное положение Wikipedia в этих историях говорит о том, что поддержка одного из самых часто посещаемых порталов обмена знаниями в мире может глубоко усилить влияние наших усилий.

Охват и качество существующего автоматического перевода. На вопрос о охвате перевода большинство носителей языков с ограниченными ресурсами выражают удовлетворение тем фактом, что их языки поддерживаются существующими системами. Несколько интервьюируемых отметили, что включение их языков в коммерчески доступные услуги делает их видимыми и повышает видимость их языков. Однако такие настроения не являются единообразными. Для отдельной группы носителей языков с ограниченными ресурсами, чьи языки содержат несколько письменных систем и вариантов, полный охват все еще отсутствует.

Например, носитель арабского языка из Марокко сказал, что для полной поддержки арабского языка необходимо учитывать различные существующие арабские языковые формы, чтобы не допустить предпочтения одной формы перед другой. Эта проблема также касается языков с двойными или множественными письменными системами (например, банджар, канури и т. д.). Исключая определенные языковые формы или письменные системы и пропагандируя более ресурсные варианты в качестве "стандартных", мы не только подвергаем сомнению точное культурное представление, но и усугубляем неравенство, которое уже присуще распределению и использованию языков по всему миру.

С другой стороны, проблемы качества вызывают обеспокоенность среди всех наших участников. Отражая ощутимый разрыв в качестве между переводом языков с высокими ресурсами и языков с ограниченными ресурсами (Джоси и др., 2019), многие интервьюируемые указывают на низкое качество и ненадежные результаты как основную причину нерегулярного или прекращенного использования. Например, носитель языка бходжпури говорит, что перевод предложения на его языке с использованием доступной коммерческой системы и последующее редактирование занимает больше времени, чем выполнение этого вручную. Другой участник заявил, что ему не нужна идеальность, а нужна технология, которая будет разумно использоваться для перевода на и с их языка. Несколько интервьюируемых даже упомянули, что недостаток внимания к их языкам в некоторых платформах перевода привел к иногда токсичным или грубым переводам, что дополнительно подрывает их уверенность в этих системах. Эти точки зрения напоминают нам, что, хотя включение языков является важным первым шагом, стремление к безопасному и высококачественному переводу все же является главным вопросом.

Кто выиграет? Обсуждения оценочной стоимости машинного перевода среди носителей языков с ограниченными ресурсами подчеркивают глубокие социоэкономические разрывы, которые разделяют одно сообщество от другого, влияя на воспринимаемую полезность технологии. В то время как машинный перевод в первую очередь помогает тем, кто из более привилегированных слоев населения, изучать новые языки или более эффективно путешествовать, его присутствие в финансово обездоленных сообществах может быть ключевым для социальной мобильности или даже экономического выживания. Например, носитель тигринского языка отмечает, что в Эфиопии, где менее 20 процентов населения имеют доступ к интернету, реальный доступ к тому, что предлагает сеть, ограничен из-за недостатка качественного перевода. Позднее он подчеркивает, что язык может быть преградой к образованию и занятости. Многие носители языков с ограниченными ресурсами из Африки повторяют эти аргументы, напоминая нам о последствиях хронической маргинализации и ее влиянии на людей (Алупо и др., 2021), и о широком спектре выгод, которые машинный перевод может принести различным населенным пунктам.

При фокусировке на отдельных общинах самих по себе видим похожие формы разделения. Например, большинство участников интервью согласны, что те, кто обладает технологическими навыками, получат больше выгоды от машинного перевода, чем те, кто этими навыками не обладает. Одно интервью намекает, что молодые люди в их общинах более подходят для использования возможностей машинного перевода, чем их старшие собратья. Указывая на пример недавней пандемии COVID-19, она отметила, что в местах, где информация, подтвержденная наукой, была ограничена из-за отсутствия доверенных формальных учреждений, старшее поколение этих общин зависело от своих более технологически подкованных сетевых и членов семьи, чтобы получать своевременную, переведенную информацию о здоровье, производную от международных организаций. В том же духе, те, кто обладает более высоким уровнем технологических навыков, также смогут лучше отражать дезинформацию, фейковые новости или интернет-мошенничество, которые могут возникнуть в результате расширения технологии перевода на языки с ограниченными ресурсами.

В целом важно отметить, что общины с языками с ограниченными ресурсами не являются монолитными группами; каждая из них навигирует уникальными социополитическими и культурными контекстами. В общении с их представителями мы узнаем, что обеспечение качественного перевода, хотя и важно по многим причинам, остается одним из решений масштабной головоломки, которая представляет собой справедливое представление языков и равный доступ к знаниям. Тем не менее, предлагая одно решение, мы надеемся вдохновить других акторов на действия. Как выразил один из говорящих на языках с ограниченными ресурсами, включение большего числа таких языков в машинный перевод помогает недооцененным языкам получить цифровую видимость на глобальном уровне, что может заставить местные институты относиться к местным языкам более серьезно и вкладывать больше ресурсов в их сохранение или обучение. Эта перспектива подчеркивает, как символические, так и материальные выгоды, которые может принести машинный перевод. Положительные поощрения со стороны говорящих на языках с ограниченными ресурсами на протяжении всего исследования напоминают нам о том, что, принимая человекоцентрический подход и сосредотачиваясь на языках, которые исторически оставались позади, мы можем помочь обществам поддерживать связь с их родными языками — неотъемлемой частью культуры и идентичности многих людей.

2.2 Нет оставленных языков: Основные принципы

Сочетая уроки, вынесенные из интервью с говорящими на языках с ограниченными ресурсами, и лучшие практики, выведенные из литературы по ответственному искусственному интеллекту (Арриета и др., 2020; Бендер и др., 2021; Блоджетт и др., 2022; Паульяд и др., 2021; Самбасиван и Холбрук, 2018), мы представляем четыре ключевых принципа, лежащих в основе наших исследований:

1. Приоритизация потребностей недооцененных сообществ. Как уже упоминалось, мы ставим потребности сообществ с языками с ограниченными ресурсами в центр наших усилий. Признавая, что машинный перевод является технологическим артефактом, который исторически деприоритизировал определённые группы населения, мы используем это усилие для перераспределения власти и ресурсов в пользу недооцененных сообществ. Поднимая потребности сообществ с языками с ограниченными ресурсами, мы надеемся, что наш вклад станет частью коллективного усилия, направленного на переход в более справедливую эпоху цифрового представления.

2. Обмен через открытое совместное использование. Говорящие на языках с ограниченными ресурсами напоминают нам, что прозрачность должна быть ключевым акцентом при разработке проекта NLLB. С двойной целью содействия прозрачности и избежания дублирования усилий, мы решили с самого начала открыть исходный код NLLB. Таким образом, научное сообщество в целом может напрямую извлечь выгоду из нашего вклада. Создание NLLB с учетом открытого совместного использования также мотивирует нас быть намеренными и осознанными в нашем подходе на протяжении всего процесса разработки. Мы надеемся, что влияние нашей работы будет усилено, когда другие ученые и практики будут строить на этом усилии для развития всей области машинного перевода.

3. Междисциплинарный подход. Как выразился один из говорящих на языках с ограниченными ресурсами, машинный перевод не является просто проблемой кодирования, поскольку в своей сущности, это вопрос, касающийся человечества. Чтобы избежать "проблемы выравнивания" (Кристиан, 2020) и обеспечить, чтобы наша система работала таким образом, что она была бы чутка к ценностям и социально ответственна, наше исследовательское усилие осуществляется междисциплинарной командой с учеными из широкого спектра гуманитарных (например, философия, этика), социальных наук (например, социология, лингвистика) и технических (например, компьютерные науки, статистика) областей. Расширение разнообразия нашей команды не только расширяет наш методологический и аналитический арсенал, но и дает нам возможность использовать различные навыки для решения различных аспектов этой задачи.

4. Рефлексивность в наших усилиях. Наконец, рефлексивность мотивирует нас критически рассматривать наши собственные суждения, практики и системы убеждений на протяжении процесса создания NLLB, чтобы обеспечить, что мы снижаем предвзятости, которые часто встречаются при разработке систем искусственного интеллекта. Конкретно мы предлагаем детализированную документацию о том, как мы пришли к различным решениям ниже, чтобы позволить различным заинтересованным сторонам изучить наши намерения и мотивации. Мы признаем, что при таких амбициозных усилиях необходимо делать компромиссы, и совершенство остается недостижимым. Таким образом, мы надеемся, что наше текущее усилие стимулирует критический анализ существующих практик, что, в свою очередь, позволит нам принимать более обоснованные решения в будущих итерациях NLLB.

Теперь, когда мы описали наши мотивацию и ценности, мы переходим к следующей части истории — преодолению технических вызовов, связанных с реализацией машинного перевода для 200 языков, начиная от идентификации языка до обучающих данных, моделей и оценки. Как и в случае с любыми передовыми вмешательствами, большие проблемы требуют новаторских адаптаций. Ниже мы описываем путь, который мы преодолели для реализации технических аспектов NLLB, подробно останавливаясь на этических и социальных аспектах. Сначала давайте познакомимся с нашими кандидатами на языки.

3. Языки

Широко доступные системы машинного перевода поддерживают около 130 языков; наша цель — увеличить это число до 200. При решении, какие языки включить, мы сначала рассмотрели 101 язык, охваченный в Flores-101, наборе данных для оценки перевода, включающем в основном малоресурсные языки. Затем мы составили предварительный список из более чем 250 возможных языковых кандидатов, который в конечном итоге сократили до около 210 для финального расширения с 101 до более чем 200 языков.

Процесс создания предварительного списка был следующим. Сначала мы рассмотрели все языки, представленные в Википедии. Как отмечалось в предыдущем разделе, Википедия является ключевым местом распространения знаний для многих носителей малоресурсных языков, что делает её актуальной отправной точкой. В настоящее время Википедия поддерживает более 300 языков, целенаправленно расширяя свой контент за пределы английского (Johnson и Lescak, 2022), и новые языки могут быть добавлены в рамках процесса запроса от сообщества.

Code	Language	Script	Family	Subgrouping	🌐	Res.	Specification
ace_Arab[NEW]	**Acehnese**	Arabic	Austronesian	Malayo-Polynesian	✗	Low	North Acehnese
ace_Latn[NEW]	**Acehnese**	Latin	Austronesian	Malayo-Polynesian	✗	Low	North Acehnese
acm_Arab[NEW]	**Mesopotamian Arabic**	Arabic	Afro-Asiatic	Semitic	✗	Low	Baghdadi
acq_Arab[NEW]	**Taʿizzi-Adeni Arabic**	Arabic	Afro-Asiatic	Semitic	✗	Low	
aeb_Arab[NEW]	**Tunisian Arabic**	Arabic	Afro-Asiatic	Semitic	✗	Low	Derja
afr_Latn	**Afrikaans**	Latin	Indo-European	Germanic	🌐	High	
ajp_Arab[NEW]	**South Levantine Arabic**	Arabic	Afro-Asiatic	Semitic	✗	Low	Ammani
aka_Latn[NEW]	**Akan**	Latin	Atlantic-Congo	Kwa Volta-Congo	✗	Low	Asante
amh_Ethi	**Amharic**	Geʿez	Afro-Asiatic	Semitic	🌐	Low	Addis Ababa
apc_Arab[NEW]	**North Levantine Arabic**	Arabic	Afro-Asiatic	Semitic	✗	Low	
arb_Arab	**Modern Standard Arabic**	Arabic	Afro-Asiatic	Semitic	🌐	High	
arb_Latn[NEW]	**Modern Standard Arabic**	Latin	Afro-Asiatic	Semitic	✗	Low	
ars_Arab[NEW]	**Najdi Arabic**	Arabic	Afro-Asiatic	Semitic	✗	Low	
ary_Arab[NEW]	**Moroccan Arabic**	Arabic	Afro-Asiatic	Semitic	✗	Low	
arz_Arab[NEW]	**Egyptian Arabic**	Arabic	Afro-Asiatic	Semitic	✗	Low	
asm_Beng	**Assamese**	Bengali	Indo-European	Indo-Aryan	🌐	Low	Eastern
ast_Latn	**Asturian**	Latin	Indo-European	Italic	✗	Low	Central
awa_Deva[NEW]	**Awadhi**	Devanagari	Indo-European	Indo-Aryan	✗	Low	Ayodhya
ayr_Latn[NEW]	**Central Aymara**	Latin	Aymaran	Central Southern Aymara	🌐	Low	Aymara La Paz jilata
azb_Arab[NEW]	**South Azerbaijani**	Arabic	Turkic	Common Turkic	✗	Low	Tabrizi
azj_Latn	**North Azerbaijani**	Latin	Turkic	Common Turkic	🌐	Low	Shirvan
bak_Cyrl[NEW]	**Bashkir**	Cyrillic	Turkic	Common Turkic	🌐	Low	Literary
bam_Latn[NEW]	**Bambara**	Latin	Mande	Western Mande	🌐	Low	
ban_Latn[NEW]	**Balinese**	Latin	Austronesian	Malayo-Polynesian	✗	Low	
bel_Cyrl	**Belarusian**	Cyrillic	Indo-European	Balto-Slavic	🌐	Low	Central
bem_Latn[NEW]	**Bemba**	Latin	Atlantic-Congo	Benue-Congo	✗	Low	Central
ben_Beng	**Bengali**	Bengali	Indo-European	Indo-Aryan	🌐	High	Rarhi
bho_Deva[NEW]	**Bhojpuri**	Devanagari	Indo-European	Indo-Aryan	🌐	Low	
bjn_Arab[NEW]	**Banjar**	Arabic	Austronesian	Malayo-Polynesian	✗	Low	Banjar Kuala
bjn_Latn[NEW]	**Banjar**	Latin	Austronesian	Malayo-Polynesian	✗	Low	Banjar Kuala
bod_Tibt[NEW]	**Standard Tibetan**	Tibetan	Sino-Tibetan	Bodic	🌐	Low	Lhasa
bos_Latn	**Bosnian**	Latin	Indo-European	Balto-Slavic	🌐	High	
bug_Latn[NEW]	**Buginese**	Latin	Austronesian	Malayo-Polynesian	✗	Low	Bone
bul_Cyrl	**Bulgarian**	Cyrillic	Indo-European	Balto-Slavic	🌐	High	
cat_Latn	**Catalan**	Latin	Indo-European	Italic	🌐	High	
ceb_Latn	**Cebuano**	Latin	Austronesian	Malayo-Polynesian	🌐	Low	
ces_Latn	**Czech**	Latin	Indo-European	Balto-Slavic	🌐	High	
cjk_Latn[NEW]	**Chokwe**	Latin	Atlantic-Congo	Benue-Congo	✗	Low	
ckb_Arab	**Central Kurdish**	Arabic	Indo-European	Iranian	🌐	Low	
crh_Latn[NEW]	**Crimean Tatar**	Latin	Turkic	Common Turkic	✗	Low	
cym_Latn	**Welsh**	Latin	Indo-European	Celtic	🌐	Low	Y Wyndodeg
dan_Latn	**Danish**	Latin	Indo-European	Germanic	🌐	High	
deu_Latn	**German**	Latin	Indo-European	Germanic	🌐	High	
dik_Latn[NEW]	**Southwestern Dinka**	Latin	Nilotic	Western Nilotic	✗	Low	Rek
dyu_Latn[NEW]	**Dyula**	Latin	Mande	Western Mande	✗	Low	
dzo_Tibt[NEW]	**Dzongkha**	Tibetan	Sino-Tibetan	Bodic	✗	Low	
ell_Grek	**Greek**	Greek	Indo-European	Graeco-Phrygian	🌐	High	
eng_Latn	**English**	Latin	Indo-European	Germanic	🌐	High	
epo_Latn[NEW]	**Esperanto**	Latin	Constructed	Esperantic	🌐	Low	
est_Latn	**Estonian**	Latin	Uralic	Finnic	🌐	High	
eus_Latn[NEW]	**Basque**	Latin	Basque	–	🌐	High	
ewe_Latn[NEW]	**Ewe**	Latin	Atlantic-Congo	Kwa Volta-Congo	🌐	Low	Aŋlo
fao_Latn[NEW]	**Faroese**	Latin	Indo-European	Germanic	🌐	Low	
fij_Latn[NEW]	**Fijian**	Latin	Austronesian	Malayo-Polynesian	🌐	Low	Bau
fin_Latn	**Finnish**	Latin	Uralic	Finnic	🌐	High	
fon_Latn[NEW]	**Fon**	Latin	Atlantic-Congo	Kwa Volta-Congo	✗	Low	
fra_Latn	**French**	Latin	Indo-European	Italic	🌐	High	
fur_Latn[NEW]	**Friulian**	Latin	Indo-European	Italic	✗	Low	Central
fuv_Latn	**Nigerian Fulfulde**	Latin	Atlantic-Congo	North-Central Atlantic	✗	Low	Sokoto
gla_Latn[NEW]	**Scottish Gaelic**	Latin	Indo-European	Celtic	✗	Low	Northern Hebrides

Code	Language	Script	Family	Subgrouping	🌐	Res.	Specification
gle_Latn	**Irish**	Latin	Indo-European	Celtic	🌐	Low	
glg_Latn	**Galician**	Latin	Indo-European	Italic	🌐	Low	
grn_Latn[NEW]	**Guarani**	Latin	Tupian	Maweti-Guarani	🌐	Low	
guj_Gujr	**Gujarati**	Gujarati	Indo-European	Indo-Aryan	🌐	Low	Amdavadi/Surti
hat_Latn[NEW]	**Haitian Creole**	Latin	Indo-European	Italic	🌐	Low	
hau_Latn	**Hausa**	Latin	Afro-Asiatic	Chadic	🌐	Low	
heb_Hebr	**Hebrew**	Hebrew	Afro-Asiatic	Semitic	🌐	High	
hin_Deva	**Hindi**	Devanagari	Indo-European	Indo-Aryan	🌐	High	
hne_Deva[NEW]	**Chhattisgarhi**	Devanagari	Indo-European	Indo-Aryan	✗	Low	
hrv_Latn	**Croatian**	Latin	Indo-European	Balto-Slavic	🌐	High	
hun_Latn	**Hungarian**	Latin	Uralic	–	🌐	High	
hye_Armn	**Armenian**	Armenian	Indo-European	Armenic	🌐	Low	Yerevan
ibo_Latn	**Igbo**	Latin	Atlantic-Congo	Benue-Congo	🌐	Low	Central
ilo_Latn[NEW]	**Ilocano**	Latin	Austronesian	Malayo-Polynesian	🌐	Low	
ind_Latn	**Indonesian**	Latin	Austronesian	Malayo-Polynesian	🌐	High	
isl_Latn	**Icelandic**	Latin	Indo-European	Germanic	🌐	High	
ita_Latn	**Italian**	Latin	Indo-European	Italic	🌐	High	
jav_Latn	**Javanese**	Latin	Austronesian	Malayo-Polynesian	🌐	Low	
jpn_Jpan	**Japanese**	Japanese	Japonic	Japanesic	🌐	High	
kab_Latn[NEW]	**Kabyle**	Latin	Afro-Asiatic	Berber	✗	Low	North Eastern
kac_Latn[NEW]	**Jingpho**	Latin	Sino-Tibetan	Brahmaputran	✗	Low	
kam_Latn	**Kamba**	Latin	Atlantic-Congo	Benue-Congo	✗	Low	Machakos
kan_Knda	**Kannada**	Kannada	Dravidian	South Dravidian	🌐	Low	Central
kas_Arab[NEW]	**Kashmiri**	Arabic	Indo-European	Indo-Aryan	✗	Low	Kishtwari
kas_Deva[NEW]	**Kashmiri**	Devanagari	Indo-European	Indo-Aryan	✗	Low	Kishtwari
kat_Geor	**Georgian**	Georgian	Kartvelian	Georgian-Zan	🌐	Low	Kartlian
knc_Arab[NEW]	**Central Kanuri**	Arabic	Saharan	Western Saharan	✗	Low	Yerwa
knc_Latn[NEW]	**Central Kanuri**	Latin	Saharan	Western Saharan	✗	Low	Yerwa
kaz_Cyrl	**Kazakh**	Cyrillic	Turkic	Common Turkic	🌐	High	
kbp_Latn[NEW]	**Kabiyè**	Latin	Atlantic-Congo	North Volta-Congo	✗	Low	Kɛ̀wɛ
kea_Latn[NEW]	**Kabuverdianu**	Latin	Indo-European	Italic	✗	Low	Sotavento
khm_Khmr	**Khmer**	Khmer	Austroasiatic	Khmeric	🌐	Low	Central
kik_Latn[NEW]	**Kikuyu**	Latin	Atlantic-Congo	Benue-Congo	✗	Low	Southern
kin_Latn[NEW]	**Kinyarwanda**	Latin	Atlantic-Congo	Benue-Congo	🌐	Low	
kir_Cyrl	**Kyrgyz**	Cyrillic	Turkic	Common Turkic	🌐	Low	Northern
kmb_Latn[NEW]	**Kimbundu**	Latin	Atlantic-Congo	Benue-Congo	✗	Low	
kmr_Latn[NEW]	**Northern Kurdish**	Latin	Indo-European	Iranian	🌐	Low	
kon_Latn[NEW]	**Kikongo**	Latin	Atlantic-Congo	Benue-Congo	✗	Low	
kor_Hang	**Korean**	Hangul	Koreanic	Korean	🌐	High	
lao_Laoo	**Lao**	Lao	Tai-Kadai	Kam-Tai	🌐	Low	Vientiane
lij_Latn[NEW]	**Ligurian**	Latin	Indo-European	Italic	✗	Low	Zeneise
lim_Latn[NEW]	**Limburgish**	Latin	Indo-European	Germanic	✗	Low	Maastrichtian
lin_Latn	**Lingala**	Latin	Atlantic-Congo	Benue-Congo	🌐	Low	
lit_Latn	**Lithuanian**	Latin	Indo-European	Balto-Slavic	🌐	High	
lmo_Latn[NEW]	**Lombard**	Latin	Indo-European	Italic	✗	Low	Western
ltg_Latn[NEW]	**Latgalian**	Latin	Indo-European	Balto-Slavic	✗	Low	Central
ltz_Latn	**Luxembourgish**	Latin	Indo-European	Germanic	🌐	Low	
lua_Latn[NEW]	**Luba-Kasai**	Latin	Atlantic-Congo	Benue-Congo	✗	Low	
lug_Latn	**Ganda**	Latin	Atlantic-Congo	Benue-Congo	🌐	Low	
luo_Latn	**Luo**	Latin	Nilotic	Western Nilotic	✗	Low	
lus_Latn[NEW]	**Mizo**	Latin	Sino-Tibetan	Kuki-Chin-Naga	🌐	Low	Aizawl
lvs_Latn	**Standard Latvian**	Latin	Indo-European	Balto-Slavic	🌐	High	
mag_Deva[NEW]	**Magahi**	Devanagari	Indo-European	Indo-Aryan	✗	Low	Gaya
mai_Deva[NEW]	**Maithili**	Devanagari	Indo-European	Indo-Aryan	🌐	Low	
mal_Mlym	**Malayalam**	Malayalam	Dravidian	South Dravidian	🌐	Low	
mar_Deva	**Marathi**	Devanagari	Indo-European	Indo-Aryan	🌐	Low	Varhadi
min_Arab[NEW]	**Minangkabau**	Arabic	Austronesian	Malayo-Polynesian	✗	Low	Agam-Tanah Datar
min_Latn[NEW]	**Minangkabau**	Latin	Austronesian	Malayo-Polynesian	✗	Low	Agam-Tanah Datar
mkd_Cyrl	**Macedonian**	Cyrillic	Indo-European	Balto-Slavic	🌐	High	
plt_Latn[NEW]	**Plateau Malagasy**	Latin	Austronesian	Malayo-Polynesian	🌐	Low	Merina
mlt_Latn	**Maltese**	Latin	Afro-Asiatic	Semitic	🌐	High	

Code	Language	Script	Family	Subgrouping	🌐	Res.	Specification
mni_Beng[NEW]	**Meitei**	Bengali	Sino-Tibetan	Kuki-Chin-Naga	✗	Low	
khk_Cyrl	**Halh Mongolian**	Cyrillic	Mongolic-Khitan	Mongolic	🌐	Low	
mos_Latn[NEW]	**Mossi**	Latin	Atlantic-Congo	North Volta-Congo	✗	Low	Ouagadougou
mri_Latn	**Maori**	Latin	Austronesian	Malayo-Polynesian	🌐	Low	Waikato-Ngapuhi
mya_Mymr	**Burmese**	Myanmar	Sino-Tibetan	Burmo-Qiangic	🌐	Low	Mandalay-Yangon
nld_Latn	**Dutch**	Latin	Indo-European	Germanic	🌐	High	
nno_Latn[NEW]	**Norwegian Nynorsk**	Latin	Indo-European	Germanic	✗	Low	
nob_Latn	**Norwegian Bokmål**	Latin	Indo-European	Germanic	🌐	Low	
npi_Deva	**Nepali**	Devanagari	Indo-European	Indo-Aryan	🌐	Low	Eastern
nso_Latn	**Northern Sotho**	Latin	Atlantic-Congo	Benue-Congo	🌐	Low	
nus_Latn[NEW]	**Nuer**	Latin	Nilotic	Western Nilotic	✗	Low	
nya_Latn	**Nyanja**	Latin	Atlantic-Congo	Benue-Congo	🌐	Low	
oci_Latn	**Occitan**	Latin	Indo-European	Italic	✗	Low	
gaz_Latn[NEW]	**West Central Oromo**	Latin	Afro-Asiatic	Cushitic	🌐	Low	
ory_Orya	**Odia**	Oriya	Indo-European	Indo-Aryan	🌐	Low	Baleswari (Northern)
pag_Latn[NEW]	**Pangasinan**	Latin	Austronesian	Malayo-Polynesian	✗	Low	
pan_Guru	**Eastern Panjabi**	Gurmukhi	Indo-European	Indo-Aryan	🌐	Low	Majhi
pap_Latn[NEW]	**Papiamento**	Latin	Indo-European	Italic	✗	Low	Römer-Maduro-Jonis
pes_Arab	**Western Persian**	Arabic	Indo-European	Iranian	🌐	High	
pol_Latn	**Polish**	Latin	Indo-European	Balto-Slavic	🌐	High	
por_Latn	**Portuguese**	Latin	Indo-European	Italic	🌐	High	Brazil
prs_Arab[NEW]	**Dari**	Arabic	Indo-European	Iranian	🌐	Low	Kabuli
pbt_Arab	**Southern Pashto**	Arabic	Indo-European	Iranian	🌐	Low	Literary
quy_Latn[NEW]	**Ayacucho Quechua**	Latin	Quechuan	Chinchay	🌐	Low	Southern Quechua
ron_Latn	**Romanian**	Latin	Indo-European	Italic	🌐	High	
run_Latn[NEW]	**Rundi**	Latin	Atlantic-Congo	Benue-Congo	✗	Low	
rus_Cyrl	**Russian**	Cyrillic	Indo-European	Balto-Slavic	🌐	High	
sag_Latn[NEW]	**Sango**	Latin	Atlantic-Congo	North Volta-Congo	✗	Low	
san_Deva[NEW]	**Sanskrit**	Devanagari	Indo-European	Indo-Aryan	🌐	Low	
sat_Beng[NEW]	**Santali**	Bengali	Austroasiatic	Mundaic	✗	Low	
scn_Latn[NEW]	**Sicilian**	Latin	Indo-European	Italic	✗	Low	Literary Sicilian
shn_Mymr[NEW]	**Shan**	Myanmar	Tai-Kadai	Kam-Tai	✗	Low	
sin_Sinh[NEW]	**Sinhala**	Sinhala	Indo-European	Indo-Aryan	🌐	Low	
slk_Latn	**Slovak**	Latin	Indo-European	Balto-Slavic	🌐	High	
slv_Latn[NEW]	**Slovenian**	Latin	Indo-European	Balto-Slavic	🌐	High	
smo_Latn[NEW]	**Samoan**	Latin	Austronesian	Malayo-Polynesian	🌐	Low	
sna_Latn	**Shona**	Latin	Atlantic-Congo	Benue-Congo	🌐	Low	
snd_Arab	**Sindhi**	Arabic	Indo-European	Indo-Aryan	🌐	Low	Vicholi
som_Latn	**Somali**	Latin	Afro-Asiatic	Cushitic	🌐	Low	Nsom
sot_Latn[NEW]	**Southern Sotho**	Latin	Atlantic-Congo	Benue-Congo	🌐	High	
spa_Latn	**Spanish**	Latin	Indo-European	Italic	🌐	High	Latin American
als_Latn[NEW]	**Tosk Albanian**	Latin	Indo-European	Albanian	🌐	High	
srd_Latn[NEW]	**Sardinian**	Latin	Indo-European	Italic	✗	Low	Logudorese and Campidanese
srp_Cyrl	**Serbian**	Cyrillic	Indo-European	Balto-Slavic	🌐	Low	
ssw_Latn[NEW]	**Swati**	Latin	Atlantic-Congo	Benue-Congo	✗	Low	
sun_Latn[NEW]	**Sundanese**	Latin	Austronesian	Malayo-Polynesian	🌐	Low	
swe_Latn	**Swedish**	Latin	Indo-European	Germanic	🌐	High	
swh_Latn	**Swahili**	Latin	Atlantic-Congo	Benue-Congo	🌐	High	Kiunguja
szl_Latn[NEW]	**Silesian**	Latin	Indo-European	Balto-Slavic	✗	Low	
tam_Taml	**Tamil**	Tamil	Dravidian	South Dravidian	🌐	Low	Chennai
tat_Cyrl[NEW]	**Tatar**	Cyrillic	Turkic	Common Turkic	🌐	Low	Central and Middle
tel_Telu	**Telugu**	Telugu	Dravidian	South Dravidian	🌐	Low	Coastal
tgk_Cyrl	**Tajik**	Cyrillic	Indo-European	Iranian	🌐	Low	
tgl_Latn	**Tagalog**	Latin	Austronesian	Malayo-Polynesian	🌐	High	
tha_Thai	**Thai**	Thai	Tai-Kadai	Kam-Tai	🌐	High	
tir_Ethi[NEW]	**Tigrinya**	Ge'ez	Afro-Asiatic	Semitic	🌐	Low	
taq_Latn[NEW]	**Tamasheq**	Latin	Afro-Asiatic	Berber	✗	Low	Kal Ansar
taq_Tfng[NEW]	**Tamasheq**	Tifinagh	Afro-Asiatic	Berber	✗	Low	Kal Ansar
tpi_Latn[NEW]	**Tok Pisin**	Latin	Indo-European	Germanic	✗	Low	
tsn_Latn[NEW]	**Tswana**	Latin	Atlantic-Congo	Benue-Congo	✗	High	Sehurutshe
tso_Latn[NEW]	**Tsonga**	Latin	Atlantic-Congo	Benue-Congo	🌐	Low	

Code	Language	Script	Family	Subgrouping	🌐	Res.	Specification
tuk_Latn[NEW]	**Turkmen**	Latin	Turkic	Common Turkic	🌐	Low	Teke
tum_Latn[NEW]	**Tumbuka**	Latin	Atlantic-Congo	Benue-Congo	✗	Low	Rumphi
tur_Latn	**Turkish**	Latin	Turkic	Common Turkic	🌐	High	
twi_Latn[NEW]	**Twi**	Latin	Atlantic-Congo	Kwa Volta-Congo	🌐	Low	Akuapem
tzm_Tfng[NEW]	**Central Atlas Tamazight**	Tifinagh	Afro-Asiatic	Berber	✗	Low	
uig_Arab[NEW]	**Uyghur**	Arabic	Turkic	Common Turkic	🌐	Low	
ukr_Cyrl	**Ukrainian**	Cyrillic	Indo-European	Balto-Slavic	🌐	High	
umb_Latn	**Umbundu**	Latin	Atlantic-Congo	Benue-Congo	✗	Low	
urd_Arab	**Urdu**	Arabic	Indo-European	Indo-Aryan	🌐	Low	Lashkari
uzn_Latn	**Northern Uzbek**	Latin	Turkic	Common Turkic	🌐	High	
vec_Latn[NEW]	**Venetian**	Latin	Indo-European	Italic	✗	Low	Venice
vie_Latn	**Vietnamese**	Latin	Austroasiatic	Vietic	🌐	High	
war_Latn[NEW]	**Waray**	Latin	Austronesian	Malayo-Polynesian	✗	Low	Tacloban
wol_Latn	**Wolof**	Latin	Atlantic-Congo	North-Central Atlantic	✗	Low	Dakkar
xho_Latn	**Xhosa**	Latin	Atlantic-Congo	Benue-Congo	🌐	High	Ngqika
ydd_Hebr[NEW]	**Eastern Yiddish**	Hebrew	Indo-European	Germanic	🌐	Low	Hasidic
yor_Latn	**Yoruba**	Latin	Atlantic-Congo	Benue-Congo	🌐	Low	Ọ̀yọ́ and Ibadan
yue_Hant[NEW]	**Yue Chinese**	Han (Traditional)	Sino-Tibetan	Sinitic	🌐	Low	
zho_Hans	**Chinese**	Han (Simplified)	Sino-Tibetan	Sinitic	🌐	High	
zho_Hant	**Chinese**	Han (Traditional)	Sino-Tibetan	Sinitic	🌐	High	
zsm_Latn	**Standard Malay**	Latin	Austronesian	Malayo-Polynesian	🌐	High	
zul_Latn	**Zulu**	Latin	Atlantic-Congo	Benue-Congo	🌐	High	

Таблица 1: 204 языка программы "No Language Left Behind": В таблице указаны код языка, название языка, письменность и языковая семья. Символ 🌐 указывает на поддержку машинного перевода Google и/или Microsoft, тогда как 7 указывает на отсутствие поддержки ни тем, ни другим. Res. указывает, классифицируем ли мы язык как высокоресурсный или малоресурсный. В спецификации, если доступно, содержится дополнительная информация о варианте языка, собранном в Flores-200. Надстрочный индекс new указывает на новые языки, добавленные в Flores-200 по сравнению с Flores-101.

Затем мы собрали списки языков, на которых говорят носители в различных регионах, с особым акцентом на африканские языки — категорию языков, которые исторически были недопредставлены в переводческих усилиях (Nekoto et al., 2020). Далее мы рассмотрели охват языков в нескольких существующих наборах данных в сообществе обработки естественного языка, уделяя особое внимание учебным наборам данных без сопутствующих наборов для оценки. Наконец, мы учитывали распространение и использование каждого языка, оценивая приблизительное количество носителей языка и другие переменные на уровне сообществ, которые были важны для нашей работы.

Затем, для каждого из языковых кандидатов, мы сотрудничали с лингвистами из различных специализированных языковых сервисов, чтобы понять, имеет ли каждый из этих языков стандартизированную письменную форму. Мы делали это потому, что наличие надежного, качественного оценочного набора данных критично для ускоренного экспериментального прогресса. Однако приоритизация языков с достаточно стандартизированными письменными формами имеет свои недостатки (см. Приложение A). Во-первых, многие языки имеют естественные вариации и пишутся по разным стандартам или сценариям в разных регионах. Например, такие языки, как фула, включают несколько различных разновидностей, а такие языки, как кашмирский и центральный канури, содержат несколько сценариев, которые часто используются. Систематическое документирование этих аспектов помогло нам оценить, как мы можем лучше поддерживать различные варианты разных языков (например, языки с несколькими системами письма или естественными вариациями).

Вместе с этими соображениями, решение о включении языков в окончательный список в конечном итоге сводилось к оценке потенциального воздействия, которое мы могли бы оказать на соответствующие сообщества с низкими ресурсами. Например, мы исключаем языки с чрезвычайно низким числом носителей. Без целенаправленного плана для тщательного понимания нужд этих сообществ и потенциальных рисков, которые мы можем причинить, мы не чувствуем себя комфортно, включая их языки в наши усилия. В соответствии с нашими руководящими принципами, многие из языков, которые вошли в окончательный список, имеют присутствие на Википедии и происходят из исторически недопредставленных регионов. Последнее, но не менее важное: стоит отметить, что в этой работе мы исключаем многие языки, не имеющие письменных стандартов или являющиеся преимущественно устными. Мы надеемся, что в будущем исследования смогут уделить больше внимания языкам с различными модальностями.

Информация о языках В соответствии с правилом Бендера (**BenderRule**) (Bender, 2019), мы суммируем информацию о каждом из наших 204 поддерживаемых языков в Таблице 1.

Код. Мы представляем каждый язык с помощью тега последовательности BCP 47, используя трехбуквенный код ISO 639-3 в качестве базового подтега, который мы дополняем подтегами скриптов ISO 15924, поскольку мы собрали ресурсы для нескольких языков на более чем одном скрипте.

Язык. Существует несколько способов наименования одного и того же языка; из-за ограничений форматирования отображается только одна из версий. Названия языков были перекрестно проверены с основными платформами лингвистической информации, такими как Ethnologue (Lewis, 2009) и Glottolog (Hammarström et al., 2022).

Скрипт. Указано английское название скрипта. Поскольку некоторые языки пишутся на нескольких скриптах, мы стремимся поддерживать это естественное разнообразие.

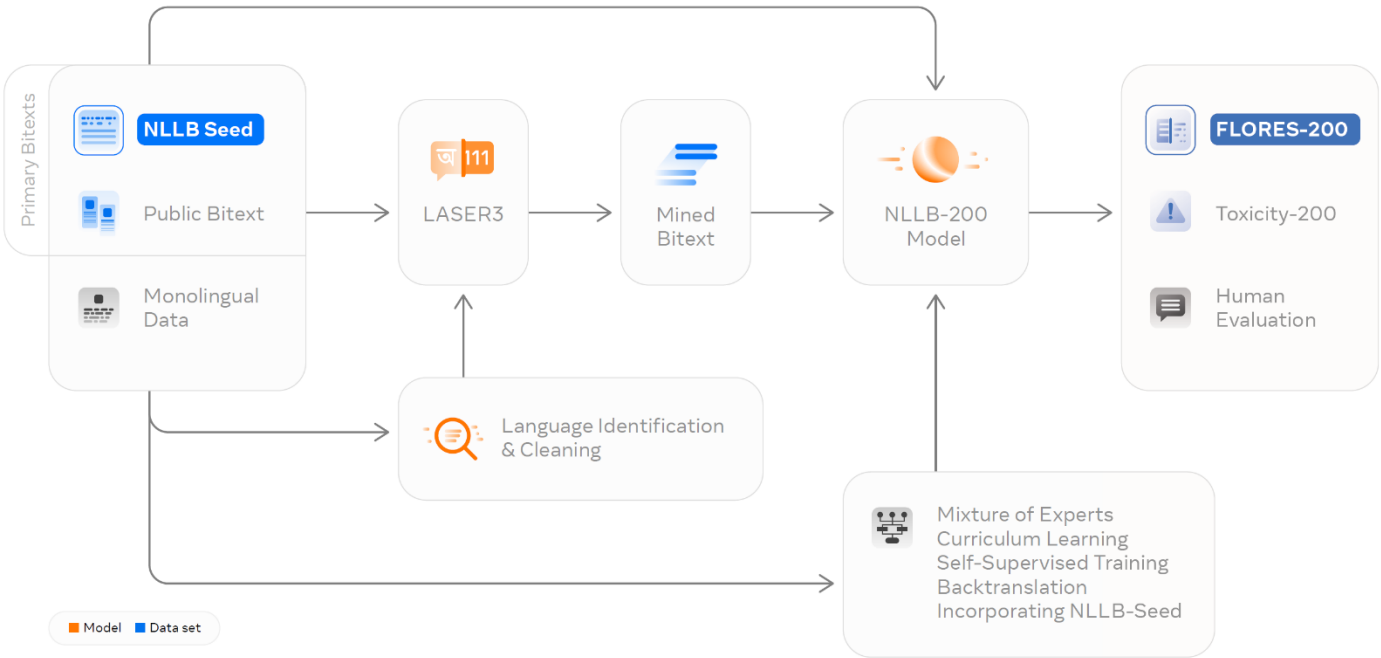
Семейство и подгруппа. Мы предоставляем информацию о языковом семействе для каждого языка на основе базы данных Glottolog (Hammarström et al., 2022).

Поддержка в Интернете. Мы проверяем, поддерживается ли каждый язык Google Translate и/или Microsoft Translate. Символ 🌐 указывает на то, что одна или обе платформы поддерживают язык. Символ 7 указывает на то, что ни одна из платформ не поддерживает язык.

Уровень ресурсов (Res). Мы классифицируем язык как с низкими ресурсами, если имеется менее 1 млн общедоступных, дублированных образцов двуязычного текста с любым другим языком в нашем наборе из 200 языков. Обратите внимание, что это выходит за рамки подсчета данных для обучения, ориентированных на английский язык, так как многие языки могут иметь доступные наборы данных на языках, более распространенных в их регионе. Например, многие страны Африки являются франкоязычными. Это приводит к тому, что 150 языков классифицируются как с низкими ресурсами.

Спецификация. Этот столбец содержит, если доступно, дополнительную информацию о конкретном варианте или регионе языка.

Информация о языках, предоставленная в Таблице 1, отражает ресурсы, собранные в рамках усилий по сбору данных Flores-200, которые описаны в следующем разделе.



4. Создание профессионально переведенных наборов данных: FLORES-200 и NLLB-Seed

Перевод для языков с низкими ресурсами сталкивается с несколькими проблемами, в первую очередь — нехваткой данных. В этом разделе мы опишем три компонента, разработанные для преодоления этой проблемы, показанные на Рисунке 3.

Во-первых, мы описываем создание Flores-200, высококачественного, многоязычного эталонного набора данных, который удваивает охват языков по сравнению с предыдущей работой, известной как Flores-101. Затем мы прослеживаем процесс разработки профессионально переведенных исходных битекстовых данных на 39 языках с низкими ресурсами, что дает нам возможность обучать любые модели, требующие параллельных данных. Наконец, мы описываем NLLB-MD, набор данных из различных доменов для оценки способности к обобщенному переводу. Эти ресурсы позволяют оценивать и создавать модели для языков, которые ранее имели минимальную поддержку.

4.1 FLORES-200

Одной из ключевых областей в исследовании машинного перевода является разработка высококачественных оценочных наборов данных или эталонов, которые можно надежно использовать для оценки прогресса в этой области. Способность оценивать позволяет нам сравнивать различные подходы и понимать, что требует дальнейших исследований и разработок. Создание эталонных наборов данных на ежегодном семинаре по машинному переводу (Akhbardeh et al., 2021) привело к быстрому прогрессу в таких направлениях перевода, как с английского на немецкий и с английского на французский. Мы также видим недавние работы по созданию наборов данных для перевода с низкими ресурсами, как это иллюстрируют наборы данных SALT (Akera et al., 2022; Babirye et al., 2022) и AmericasNLI (Ebrahimi et al., 2022). За пределами области перевода оценочные эталоны, такие как SQuAD (Rajpurkar et al., 2016), GLUE (Wang et al., 2018) и даже эталон Penn Treebank для языкового моделирования (Mikolov and Zweig, 2012), способствовали значительному прогрессу в исследованиях.

Создание Flores-200 направлено на удвоение существующего охвата языков по сравнению с Flores-101. Это вызывает значительные трудности из-за еще меньшего количества ресурсов для языков, которые мы представили в этой работе. В частности, эти языки могут требовать все более специализированных профессиональных переводчиков, имеют меньше стандартов, и процесс проверки для обеспечения качества перевода становится более сложным. Ниже, после краткого обзора характеристик Flores-101, мы подробно опишем, как мы преодолеваем эти новые трудности в создании Flores-200, уделяя особое внимание адаптированному протоколу и механизмам обеспечения качества. Затем мы представим анализ общей оценки качества нашего эталонного набора данных.

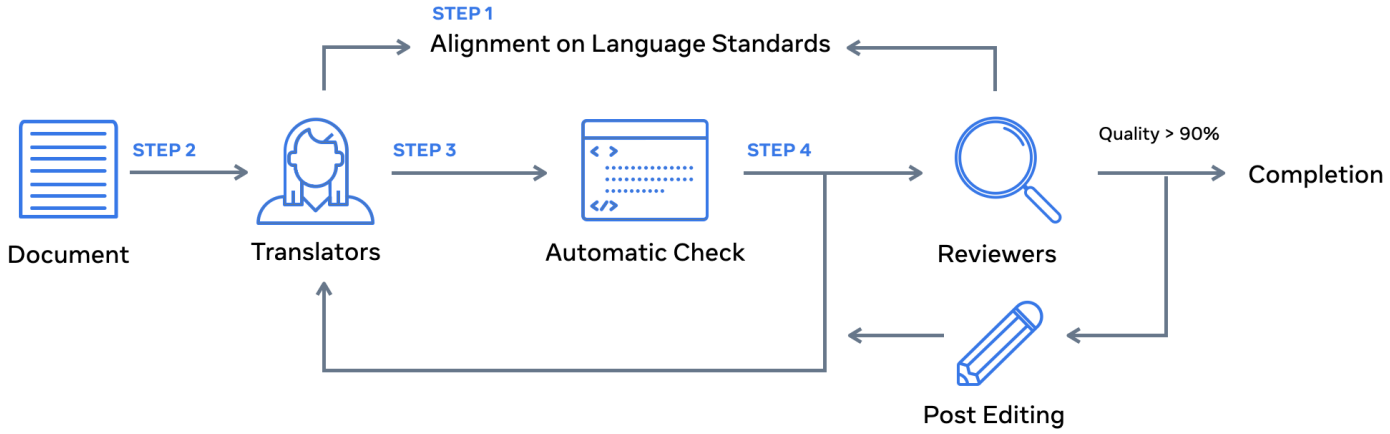


Рисунок 4: **Рабочий процесс перевода FLORES-200** Мы создали сложный многоэтапный процесс для обеспечения качества. Сначала профессиональные переводчики и рецензенты согласовывали языковые стандарты. Затем переводчики переводили полный набор предложений Flores-200, после чего проводились автоматические проверки. После этого группа независимых рецензентов проверяла качество, и на основе их оценки некоторые переводы отправлялись на пост-редактирование. Если оценка качества показывала, что качество перевода выше 90 процентов, язык считался готовым для включения в Flores-200.

4.1.1 Создание эталона для языков с низкими ресурсами

Предварительные замечания.** Как значительное расширение проекта Flores-101, проект Flores-200 состоит из 3001 предложения, отобранного из англоязычных проектов Wikimedia для **204** языков. Примерно одна треть предложений была собрана из каждого из следующих источников: Wikinews, Wikijunior и Wikivoyage. Содержание было профессионально переведено на более чем 200 языков для создания Flores-200. Поскольку мы переводим один и тот же набор контента на все языки, Flores-200 является многоязычным эталоном для перевода "многие ко многим". Подробная информация представлена в работе Goyal et al. (2022).

Поиск профессиональных переводчиков и рецензентов переводов Flores-200 создается с участием профессиональных переводчиков, которые переводят исходный набор данных FLORES на целевые языки, и отдельной группы независимых рецензентов переводов, которые проводят оценку качества человеческих переводов и предоставляют обратную связь переводчикам. И переводчики, и рецензенты проходят проверку, организованную провайдерами языковых услуг (LSP). Переводчики должны быть носителями целевого языка, иметь образование на целевом языке и высокий уровень владения английским. Переводчики должны иметь как минимум два-три года опыта перевода для соответствующей языковой пары, если у них есть академическая степень в области перевода или лингвистики, и три-пять лет опыта перевода, если у них нет соответствующей академической квалификации. Переводчики также проходят тест на перевод каждые 18 месяцев для оценки их качества перевода.

Кроме того, рецензенты Flores-200 также должны быть носителями целевого языка. Рецензенты обычно имеют степень в области перевода, как минимум пять лет опыта работы переводчиком, опыт рецензирования переводов и, если возможно, аккредитацию от соответствующего совета по переводу.

Мы отмечаем, что это строгие стандарты, и расширение Flores-200 до еще большего числа языков с низкими ресурсами в будущем может быть затруднено. Уже для многих языков найти рецензента, соответствующего указанным критериям, очень сложно. В таких случаях мы модифицировали процесс квалификации, принимая заявки от рецензентов с более общими языковыми степенями, такими как лингвистика или африканские языковые исследования, или без степени, при условии, что у них есть обширный коммерческий опыт перевода (например, более 10 лет). Для охвата еще большего числа языков с низкими ресурсами в будущем мы полагаем, что существует несколько способов работы с опытными и квалифицированными переводчиками, сохраняя при этом высокие стандарты качества. Например, одним из таких решений является перевод с неанглоязычных исходных языков. Мы проводим пилотный процесс и подробно описываем его в разделе 4.1.2.

Рабочий процесс перевода Flores-200 Рабочий процесс создания данных Flores-200 включает оригинальные процессы Flores-101, а также несколько новых начальных этапов, как подробно описано ниже.

Фаза выравнивания Мы ввели начальную фазу выравнивания в рабочий процесс для переводчиков и рецензентов перед переводом Flores-200. Включены несколько шагов по выравниванию между агентствами по переводу и обеспечению качества: выравнивание по ресурсам и целевым регионам, лингвистическим темам между переводчиками и рецензентами для каждого языка через новый шаблон выравнивания и журналов запросов между лингвистами с обеих сторон. Шаблон выравнивания помог лингвистам определить подходы к языковому письму, стандартизации, орфографии, заимствованным терминам, неологизмам, стилю информативного контента, ресурсам для глоссариев и образцам контента на целевом языке. Это было особенно полезно для языков с менее устоявшимися стандартами перевода.

Фаза перевода Затем перевод начинается с начальной фазы перевода, где все участвующие переводчики переводят одни и те же 200 предложений для каждого языка. Начальные данные для перевода равномерно распределены между тремя источниками — Wikinews, Wikijunior и Wikivoyage, при этом сегменты соответствуют одним и тем же статьям для контекста и непрерывности. Начальные переводы затем отправляются команде QA LSP для проверки. Основное внимание на начальном этапе перевода и QA уделяется пониманию и согласованию подхода к переводу между переводчиками и рецензентами. Отчет содержит обратную связь на уровне предложений (категория ошибки, уровень серьезности ошибки и комментарии, где это возможно) и высокоуровневую обратную связь по ожиданиям локали, использованию указанного письма, использованию заимствований и/или неологизмов, именованным сущностям, общему стилю и регистру.

Итерация Команды LSP по переводу могут ответить на начальные отчеты QA с арбитражем. Затем вносятся корректировки во все материалы по выравниванию, где это необходимо, и подход к переводу обновляется и повторно выравнивается. Затем начинается полный перевод всех 3000 предложений (см. Goyal et al. (2022) для подробностей).

Завершение После завершения полного перевода команда QA LSP проводит окончательную проверку QA и оценивает 20% выборку данных. В случае если окончательный балл качества перевода ниже 90%, могут следовать необязательные арбитраж, доработка и выборочные проверки QA.

4.1.2 Создание эталона для направлений, не связанных с английским языком

Стандартный рабочий процесс Flores-200 сосредоточен только на переводе с английского. Хотя это стандартизирует процесс для всех языков, у него есть очевидные ограничения. Например, есть много квалифицированных переводчиков, которые могут не говорить по-английски, но способны переводить между несколькими неанглийскими языками. Кроме того, для некоторых языков может быть проще переводить с неанглийского источника. Вместо этого мы сосредотачиваемся на адаптации и транслитерации и разрабатываем индивидуальные рабочие процессы QA для поддержки этого.

Перевод арабских языков Мы применяем этот рабочий процесс для создания наборов данных для различных вариантов арабского языка, расширяя наше языковое покрытие за пределы современного стандартного арабского языка до региональных вариантов, таких как марокканский арабский. Для создания Flores-200 для арабских вариантов команды LSP проанализировали лингвистические характеристики каждого арабского языкового варианта и степень их отличий от современного стандартного арабского по различным лингвистическим аспектам, таким как различия в словарном составе, грамматические и структурные различия, влияние других региональных языков и стиль информативного контента. На основе этих анализов арабские языковые варианты переводились либо непосредственно с английского, либо адаптировались из набора данных современного стандартного арабского с предоставлением английского источника в качестве контекста. Для каждого языкового варианта, который реализовал адаптацию, команды LSP также создавали список терминов, состоящий из терминов современного стандартного арабского и эквивалентного термина в целевом арабском варианте, чтобы обеспечить последовательную адаптацию.

Были созданы два уровня оценки качества для адаптации из современного стандартного арабского. Один уровень включал частичную проверку качества, где рецензент оценивал 10% выборку данных и проверял список терминов. Этот процесс применялся к языковым вариантам, которые, по оценкам, имели в основном различия в словарном составе, некоторые структурные различия и некоторое влияние других региональных языков. Другой уровень требовал от рецензента только оценки списка терминов, так как языковые варианты отличались от современного стандартного арабского минимально и по использованию словарного запаса. Порог качества в 90% применяется, как обычно.

Транслитерация скрипта

Было четыре языка (ace_Arab, bjn_Arab, min_Arab, taq_Tfng), которые были транслитерированы из их латинских эквивалентов. Переводческая LSP выполняет транслитерацию в соответствующие скрипты. QA LSP проверяет 20% выборку транслитерированного текста с предоставлением английского источника и данных на латинице для контекста. В отчете QA ошибки транслитерации помечаются только по уровню серьезности; категорий ошибок для транслитерации нет. Две или более ошибок, найденных в одном сегменте, помечаются как серьезные. Любое меньшее количество помечается как незначительные ошибки. Порог качества для транслитерации составляет 95%.

Overview Statistics	
# of sentences	3001
Avg # of words/sentence	21
# of articles	842
Split	**# of sentences**
dev	997
devtest	1012
test	992

# of Languages requiring Re-translation	10
Avg # of Re-translations	1
Max # of Re-translations	2
Avg # of Days to Translate 1 language	42
Avg # of Days to align	28
Avg # of Days for 1 language	119
Shortest Turnaround (days) for 1 language	70
Longest Turnaround (days) for 1 language	287

Таблица 2: **Обзор FLORES. (слева)** FLORES разделен на три оценочные части, всего 3001 предложение. **(справа)** Сводка контроля качества на основе статистики по 73 языкам, которые внедрили новый рабочий процесс Flores-200.

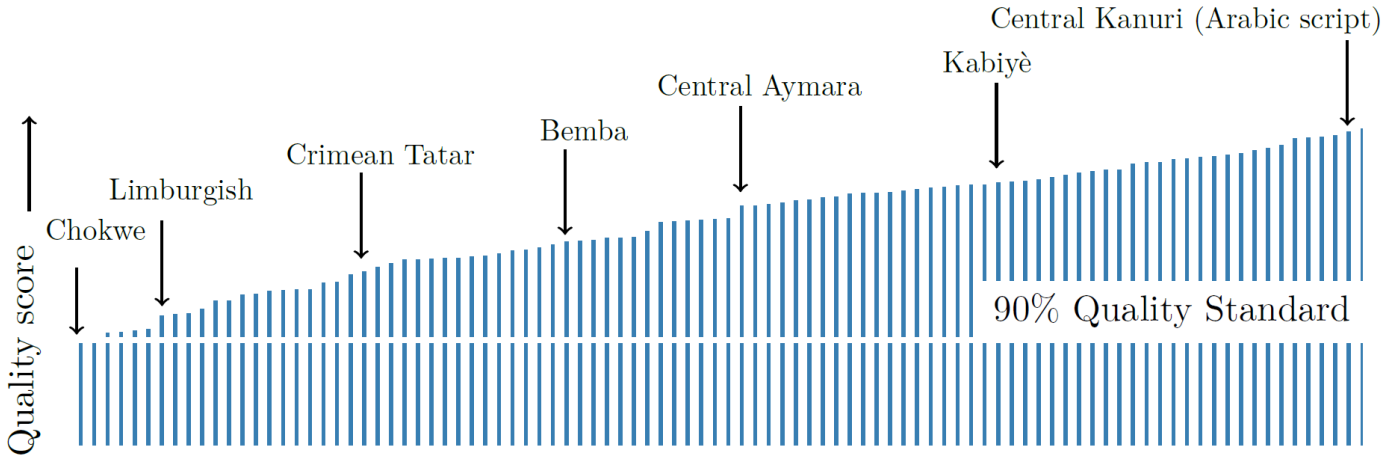


Рисунок 5: **Качество FLORES-200**: Мы отображаем оценку контроля качества для языков в Flores-200. Минимально допустимый стандарт составляет 90 процентов.

4.1.3 Краткий обзор Flores-200

Обзор.** Flores-200 состоит из переводов 842 уникальных веб-статей, в общей сложности 3001 предложение. Эти предложения разделены на три набора: dev, devtest и test. Мы публикуем полный текст наборов dev и devtest, а тестовый набор остаётся скрытым через сервер оценки. В среднем предложения содержат примерно 21 слово. Мы резюмируем эту информацию в Таблице 2 (слева).

Качество.** Для того чтобы язык считался готовым для включения в Flores-200, требуется финальная оценка качества, проведенная человеком. Мы отображаем оценки качества наших языков на Рисунке 5 с указанием нескольких примеров языков. Наиболее распространёнными ошибками при оценке качества человеческих переводов остаются ошибки в переводе и неестественные переводы. Это связано в основном с влиянием других неанглийских языков, которые могут быть широко использованы в целевых сообществах, что приводит к чрезмерным заимствованиям лексики и грамматики, буквальным переводам из-за редкого использования целевого языка в официальном, информативном стиле и более низкому уровню стандартизации. Также наблюдается растущая тенденция к орфографическим несоответствиям в человеческих переводах из-за низкого уровня стандартизации, что приводит к непоследовательным или даже субъективным предпочтениям.

Проблемы при создании наборов данных для языков с очень ограниченными ресурсами В целом, по сравнению с Flores-101, наш новый процесс перевода существенно упрощает переводческие усилия. Например, количество языков, требующих повторного перевода (см. Таблицу 2, справа), составляет всего 10, по сравнению с 45 в Flores-101. Однако, несмотря на эти улучшения, мы продолжаем сталкиваться с аналогичными проблемами, как и в Flores-101, но в ещё большем масштабе из-за растущего ограничения ресурсов для языков. Например, языки с ограниченными ресурсами не так часто используются в областях локализации или перевода. В результате, уровень стандартизации в отрасли ниже, что приводит к более сложному пути (Skadiņš et al., 2014a). Это привело к более длительным срокам выполнения, и часто требовалось несколько раз находить новых переводчиков и рецензентов. Эти проблемы особенно ощущались при работе с более сложными языками, такими как сицилийский и бугийский, на завершение которых потребовалось значительно больше времени (287 дней).

4.2 Набор данных NLLB Seed

Машинное обучение известно своей потребностью в больших объемах данных, что приводит к множеству областей исследований, направленных на снижение требуемого уровня надзора. Недавние достижения в обучении с нулевым участием человека (Chen et al., 2021; Gu et al., 2019; Johnson et al., 2017; Zhang et al., 2020) и самообучении (Bapna et al., 2022; Liu et al., 2020; Ma et al., 2021), например, стремятся уменьшить эту зависимость. Однако задачи генерации, такие как перевод, вероятно, не смогут достичь желаемого уровня качества без начальных данных. Например, сложно создать хороший перевод, не увидев минимального количества предложений на новом языке. Точно так же может быть трудно классифицировать, на каком языке написано предложение, не увидев надежных примеров текста на разных языках. Для этой цели мы создаем NLLB-Seed, набор профессионально переведенных предложений в домене Википедии. NLLB-Seed состоит из около шести тысяч предложений на 39 языках.

Такой набор данных имеет множество потенциальных применений. Критически важно, что NLLB-Seed содержит данные, которые однозначно принадлежат указанному языку, поскольку они полностью переведены профессионалами. Данные на целевом языке из NLLB-Seed можно использовать для моделей идентификации языка, которые классифицируют, на каком языке написан произвольный текст. Набор данных также можно использовать для выровненного битекста, например, для обучения моделей перевода. Другим вариантом является использование NLLB-Seed для настройки модели под определённый домен, например, адаптации универсальных моделей перевода к домену Википедии.

Выбор исходных предложений Данные для NLLB-Seed были выбраны из "Списка статей, которые должны быть в каждой Википедии" Wikimedia, который состоит из 10 000 Wikidata ID, соответствующих значимым темам в различных областях знаний и человеческой деятельности. Эти темы разделены на 11 категорий, таких как Люди, История, Философия и Религия, География. Мы равномерно отобрали подмножество ID, из которого будем выбирать данные, и сопоставили их с соответствующими статьями английской Википедии. Из каждой из этих статей мы затем выбрали данные, которые будут отправлены переводчикам. Вместо

извлечения отдельных предложений, что оставило бы переводчиков без контекста, мы решили выбирать триплеты смежных предложений, обеспечивая использование не более одного триплета из каждой статьи (аналогично Flores-200).

Следует отметить, что, как и в случае с Flores-200, исходные данные NLLB-Seed ориентированы на английский язык и выбраны из английской Википедии. Это оказывает важное влияние: содержание отражает то, что редакторы Википедии считают важным для английской Википедии, и, вероятно, не охватывает разнообразное содержание из разных культур. Кроме того, целевой текст в NLLB-Seed в конечном итоге переводится людьми, что потенциально может приводить к эффектам переводческого языка (часто определяемого как неловкие, неестественные или слишком буквальные переводы) (Volansky et al., 2015).

Процесс перевода Сценарий, спецификация, правописание и подходы к переводу были впервые установлены и согласованы на основе Flores-200. Переводчики ссылались на эти лингвистические согласования при работе над переводом данных для NLLB-Seed. Датасеты были переведены напрямую с английского на 39 языков, в то время как два языка с арабским скриптом (Ачехский и Банджар) и тамашек в скрипте Тифинаг были транслитерированы из соответствующих датасетов латинского скрипта, которые сначала были переведены с английского. После фазы перевода или транслитерации последовала фаза лингвистической оценки качества, в ходе которой завершенные датасеты проверялись на соответствие лингвистическим согласованиям из FLORES, а также на автоматические проверки качества. Датасеты затем были окончательно завершены и закончены.

Отмечаем, что NLLB-Seed имеет ключевое отличие по сравнению с оценочными бенчмарками, такими как Flores-200. Критически важно, что NLLB-Seed предназначен для использования в обучении, а не для оценки моделей. Из-за этого различия NLLB-Seed не проходит через процесс человеческой оценки качества, присутствующий в Flores-200.

4.3 Многодоменный датасет NLLB

Избежание переобучения и достижение высокой производительности на данных из различных доменов остаются важными задачами в нейронном машинном переводе (Кен и Ноулз, 2017). Хотя как Flores-200, так и NLLB-Seed охватывают множество тематик, мы хотим обеспечить хорошее качество работы моделей на текстах из разных доменов. Кроме того, так как потенциальные пользователи могут быть заинтересованы в настройке общих моделей перевода под конкретные приложения, мы хотим исследовать, насколько эффективно наша система может быть донастроена на датасете, охватывающем новый домен. Конкретно мы хотим ответить на следующие два вопроса: (1) Насколько хорошо модели обобщаются на домены, не связанные с Wikimedia? (2) Приводит ли донастройка на высококачественных параллельных текстах внутри домена к хорошей производительности? Для изучения этих вопросов мы создаем многодоменный параллельный датасет NLLB-MD, охватывающий шесть направлений и состоящий из 3000 профессионально переведенных предложений в каждом из четырех различных доменов.

Выбор языков. NLLB-MD включает следующие шесть языков: Центральный аймара (ayr_Latn), Бходжпури (bho_Deva), Дюла (dyu_Latn), Фриульский (fur_Latn), Русский (rus_Cyrl) и Волоф (wol_Latn). Вместе с пятью языками с низкими ресурсами мы также выбрали один язык с высокими ресурсами для возможности сравнения с другими моделями и датасетами. Мы выбрали языки с низкими ресурсами, связанные с другими языками с высокими ресурсами (например, fur_Latn связан с ita_Latn), чтобы облегчить будущие исследования в области языкового трансфера.

Выбор доменов. Мы собрали по 3000 английских предложений в каждом из четырех различных доменов и отправили их профессиональным переводчикам для перевода на шесть целевых языков NLLB-MD. Рабочий процесс перевода аналогичен тому, что использовался для NLLB-Seed. Включенные домены следующие:

• **Новости**: Мы перевели английскую часть разработочного набора WMT21 English-German, содержащего выборку газетных статей 2020 года (Akhbardeh и др., 2021).

• **Сценаризированная формальная речь**: Мы перевели текст, извлеченный из серии сценаризированных англоязычных выступлений по различным темам.

• **Несценаризированная неформальная речь**: Мы извлекли 3000 высказываний из мульти-сессионного чат-датасета Xu и др. (2022), содержащего в среднем 23 слова на каждый поворот.

• **Здравоохранение**: Мы перевели один отчет Всемирной организации здравоохранения (Donaldson и Раттер, 2017) и объединили его с предложениями, переведенными из английской части отчета TAUS Corona Crisis Report.

4.4 Заключение

Подводя итог, Flores-200, которая позволяет надежно оценивать более 200 языков, играет критическую роль в обеспечении качества результатов, генерируемых нашими системами. NLLB-Seed играет важную роль в обучении как представителей предложений (см. раздел 5), так и моделей перевода (см. раздел 6.5). Наконец, мы используем NLLB-MD для измерения обобщаемости наших моделей перевода в различных доменах (см. раздел 8.3). Теперь, когда мы описали создание трех датасетов с человеческим переводом и их применение, рассмотрим, как мы получили обучающие данные для нашей работы в следующем разделе.

5. Автоматическое создание обучающих данных для перевода на сотни языков

Существующие техники обучения моделей перевода сложно применить в условиях ограниченных ресурсов — то есть, когда данные для языка ограничены как в алайненных текстовых данных (битекст, или пары переведенных предложений), так и в данных на одном языке (одноязычные данные, или данные только на одном языке). Фактически, многие языки с ограниченными ресурсами поддерживаются только через небольшие целевые наборы битекстов, такие как христианская Библия (Маккарти и др., 2020), которые

чрезвычайно ограничены в разнообразии доменов. В этом разделе мы подробно описываем, как мы создали масштабный набор данных, охватывающий сотни языков, и обсуждаем вызовы, с которыми мы столкнулись из-за шумных данных в масштабах веба.

Для контекста, общедоступные данные битекста часто являются редкостью (Гауда и др., 2021). Наш подход основан на расширении существующих наборов данных путем сбора несоотнесенных одноязычных данных и использования масштабного добычи данных (Швенк и др., 2021b) для идентификации предложений с высокой вероятностью перевода друг на друга на различных языках. Для обеспечения этого для сотен языков мы сначала разрабатываем системы идентификации языка (LID, раздел 5.1), которые определяют на каком языке написан данный текст. Затем мы отбираем доступные одноязычные данные, применяем разделение на предложения и LID вместе с различными фильтрующими механизмами (раздел 5.2), и затем переходим к добыче соотнесенных пар (раздел 5.3). Обзор этого процесса представлен на рисунке 7.

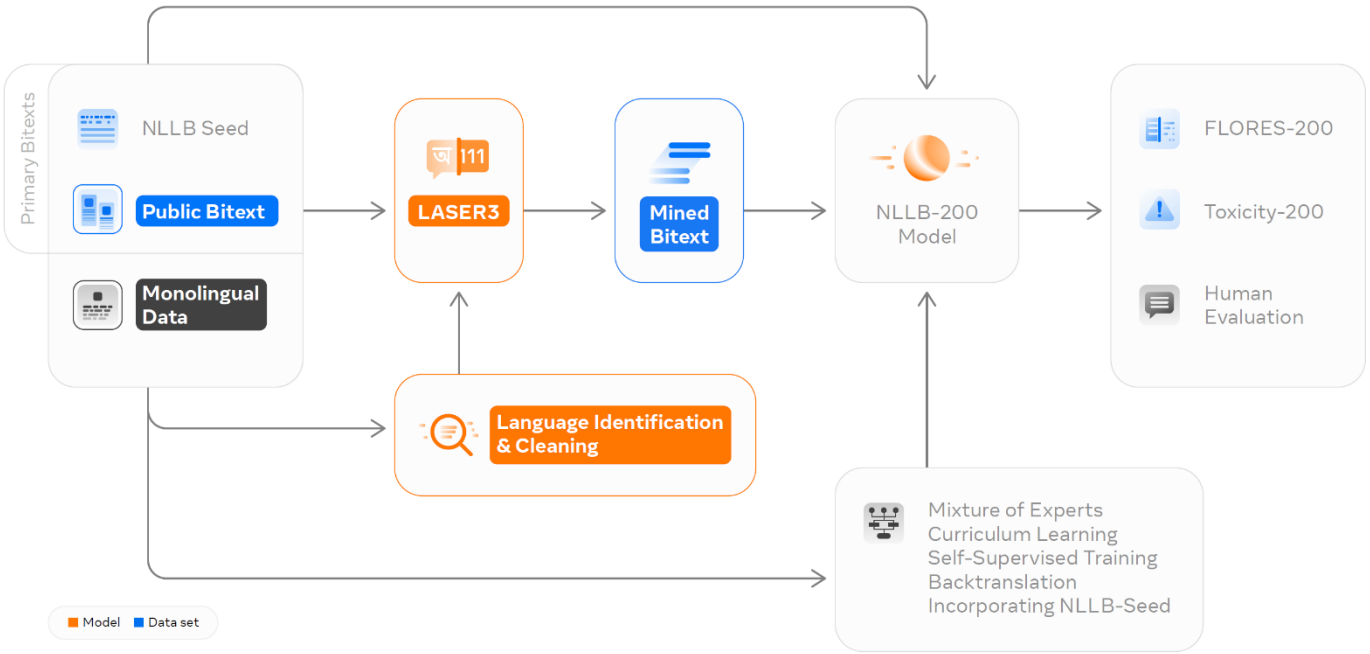


Рисунок 6**: Вклад автоматического создания набора данных проекта No Language Left Behind**: Как показано, мы создаем процесс идентификации языка и очистки одноязычных данных, затем описываем обучение LASER3 для создания масштабированных добытых битекстов для сотен языков.

5.1 Идентификация языка

Идентификация языка (LID) — это задача предсказания основного языка для определенного текстового фрагмента. Она широко используется в коммерческих приложениях (например, функция определения языка, встроенная в некоторые веб-браузеры), и имеет особое значение в исследованиях обработки естественного языка. Развитие крупномасштабного предварительного обучения, особенно увеличение внимания к мультиязычным моделям, сильно зависит от наличия и идентификации монолингвальных данных в масштабах. Продвижение кросс-языкового обучения представлений (Conneau и Lample, 2019; Wang и др., 2020b), таких как крупномасштабная добыча битекстов (Баньон и др., 2020; Рамеш и др., 2022; Швенк и др., 2021b), несупервизированный машинный перевод (Conneau и др., 2020; Рен и др., 2019; Янг и др., 2018) и масштабированное обратное перевод, требуют больших объемов чистых монолингвальных данных. Эти различные подходы, включая нашу работу по крупномасштабной добыче соотнесенных предложений, включают в себя обработку больших объемов входного текста, часто извлекаемого из веб-корпусов, таких как CommonCrawl13, и помечают их соответствующими языками.

Существует несколько известных вызовов, связанных с крупномасштабной и точной идентификацией языка с использованием веб-данных (Касуэлл и др., 2020): (1) Дисбаланс доменов может возникнуть из-за недостатка текстов, надежно помеченных по языкам. Например, христианская Библия была переведена на множество языков.

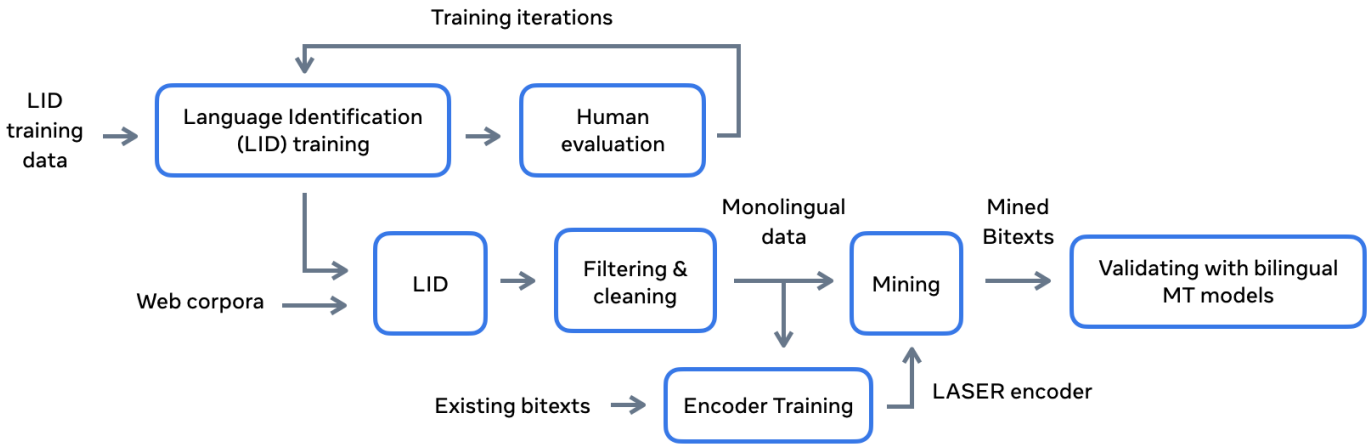


Рисунок 7: **Обзор нашего конвейера по добыче битекстов**. Идентификация языка применяется к веб-корпусам для извлечения одноязычных предложений. Затем с помощью LASER3 идентифицируются соотнесенные пары.

Однако система идентификации языка, обученная на этом корпусе, не сможет надежно классифицировать предложения из не-библейских доменов. Проблема корректного расширения обучающих данных не тривиальна: хотя веб содержит данные на тысячах языков (Прасад и др., 2018; Сканнелл, 2007), большинство из них не имеют меток. В заполнении этой пробелы помогает Википедия, которая часто используется для обучения идентификации языка (Тома, 2018) в более широком масштабе, чем христианская Библия (хотя такой относительно чистый формальный текст не является типичным представителем всемирной паутины); (2) Существует серьезный дисбаланс классов из-за того, что многие языки с ограниченными ресурсами, которые нас интересуют, слабо представлены в сети. Для работы классификаторов необходимо иметь крайне низкий уровень ложных срабатываний. В противном случае языки с ограниченными ресурсами могут быть склонны к неправильной идентификации; (3) Эффективность работы с большими веб-коллекциями остается низкой. Несмотря на то что классификация может выполняться в массовом параллелизме, работа с всеми текстами требует высокой скорости выполнения. В этом разделе мы описываем наш подход к идентификации языка и то, как мы достигаем необходимого баланса между предсказательной точностью и масштабируемостью.

5.1.1 Связанные работы

Существует обширная литература, посвященная разработке систем идентификации языка (LID). Яухиайнен и др. (2019) представляют недавний и всеобъемлющий обзор признаков и алгоритмов, используемых в литературе. В то время как в некоторых областях LID может рассматриваться как решенная проблема (МакНами, 2005), для веб-данных она остается открытой задачей (Абаджи и др., 2022; Касуэлл и др., 2020; Зампьери и др., 2015). Конкретно проблемы сосредотачиваются вокруг (1) масштабирования успешных подходов на большее количество языков (Яухиайнен и др., 2017); (2) случаев значительного несоответствия доменов (Виддоус и Брю, 2021) в случае коротких твитов или множественных языков (Дувенхаге, 2019); и (3) различия схожих языков (Гутте и др., 2016).

Масштабирование LID на большее количество языков. Внимание, уделенное развитию техник LID, привело к заметному увеличению как охвата языков, так и точности с течением времени. CLD314 и fasttext (Грейв и др., 2018) — две готовые модели, предлагающие высокую производительность определения для 107 и 187 языков соответственно. Используя многочисленные общедоступные наборы данных, (2020) и Браун (2014) сообщают о еще более широком охвате — поддержка 464 и 1366 языков соответственно. Тем не менее, разработки в области языков с ограниченными ресурсами продвигаются медленно из-за упора на религиозные тексты и ограничений, вызванных локализацией программного обеспечения. Касуэлл и др. (2020) расширяются до 1,629 языков, используя списки слов и самонадзор для создания обучающих данных, найденных в сети. Эти подходы, использующие найденные данные, страдают от дисбаланса доменов: из-за различий в доступных текстовых доменах по языкам, классификатор смешивает домен и язык. В отличие от этого, мы создаем Flores-200 в качестве развивающегося набора, чтобы производительность нашей системы LID была настроена на унифицированный микс доменов. Конечно, можно использовать христианскую Библию как унифицированный домен. Однако мы считаем, что Flores-200 ближе к контенту веба.

Несоответствие доменов. Поскольку веб охватывает очень широкий набор доменов, а надежно помеченные тексты являются редкостью, практически всегда возникает несоответствие доменов между обучающими данными и классифицируемым веб-текстом. Виддоус и Брю (2021) описывают новую особенность, основанную на ранге слов в частотных таблицах, которая повышает устойчивость систем LID к несоответствиям доменов. Они обучают свой классификатор на Википедии и сообщают о результатах на тестовом наборе Twitter, к сожалению, охватывающем только 22 языка. Короткий текст рассматривается в работе Дувенхаге (2019) для южноафриканских языков с использованием стекового классификатора. Стратегии на основе нейронных сетей также разрабатываются в работах Ансари и др. (2021); Шекхар и др. (2020) для работы с текстами, написанными на смешанных комбинациях английского и индийских языков (код-миксинг). Касуэлл и др. (2020) тщательно анализируют и классифицируют режимы отказа системы идентификации языка на веб-корпусах. Они предлагают использовать серию фильтров в сочетании с новым подходом несупервизированного обучения для радикального повышения точности при ограниченных затратах на отзыв. Однако разработка и настройка этих фильтров для всех языков затратны. Некоторые из них были успешно внедрены в практику в работе Абаджи и др. (2022) для создания более чистой версии набора данных OSCAR. Наш подход комбинирует модель fasttext (Грейв и др., 2018), обученную на наборе Flores-200, с небольшим набором ручных правил для обработки обратной связи пользователей по ошибкам классификации.

Обработка схожих языков. Выделение среди схожих языков является активной темой исследований, например, в рамках задачи различения схожих языков на конференции VarDial (Гутте и др., 2016). Несколько распространенных алгоритмов машинного обучения вместе с стандартными нейронными сетями сравниваются в работе Хааса и Дерчински (2021) для скандинавских языков. Дувенхаге (2019); Гутте и др. (2014); Зампьери и др. (2015a) исследуют различные иерархические подходы, которые сначала предсказывают

языковую группу входного текста, а затем применяют более специализированный классификатор для различия языков внутри этой группы. В данной работе мы сотрудничаем в тесном партнерстве с лингвистами для понимания, какие языки могут быть легко перепутаны, и анализируем производительность модели при использовании стратегии плоской классификации.

5.1.2 Модели

Мы используем fasttext для обучения моделей идентификации языка (Бояновский и др., 2017; Жулен и др., 2017). fasttext широко используется для задач классификации текстов благодаря своей простоте и скорости при достижении хорошего качества. Мы встраиваем символьные n-граммы из входного текста, затем применяем многоклассовый линейный классификатор сверху. Легковесная природа fasttext позволяет нашим моделям LID обрабатывать данные масштаба веба. Кроме того, линейная модель обладает преимуществом легкой интерпретируемости, что позволяет отслеживать любую ошибку классификации до её корневой причины. Это является ключевым в решении

Конструктор классификатора. Мы экспериментировали с двумя различными дизайнами. (1) Комбинация нескольких бинарных классификаторов, где окончательное решение принимается путем выбора языка с наивысшим баллом после применения порога. Мы оптимизируем порог таким образом, чтобы при низкой уверенности классификатора соответствующий язык не рассматривался для окончательного решения. Если ни один из классификаторов не превышает свой порог, предложение отфильтровывается. (2) Многоклассовый классификатор с использованием softmax для всех возможных языков. В этом случае оптимизация порога выполняется после softmax.

Наши эксперименты побудили нас сосредоточиться на втором подходе, который предлагает несколько преимуществ. Во-первых, изменение порога для одного языка не влияет на производительность других, что не является правдой для первой настройки. Во-вторых, мы обнаружили, что этот подход лучше обобщается на данные, выходящие за рамки основного использования (например, Википедия и веб-данные). Наконец, одиночный классификатор имеет дополнительное преимущество в вычислительной простоте, что упрощает процесс идентификации языка.

Обучение данных и работа с массовым дисбалансом классов. Мы используем общедоступные наборы данных для обучения нашей системы LID, частично покрывая наши интересы. Мы дополняем их NLLB-Seed (см. раздел 4.2) для любых отсутствующих языков. Однако объем данных для каждого языка далеко не одинаков, и существует массовый дисбаланс классов в исходных данных для обучения (Касуэлл и др., 2020; Данн, 2020). Например, только английский составляет 10.1% наших данных для обучения, в то время как минангкабау (латиница) представляет всего лишь 0.06%. В соответствии с работой Аривазаган и др. (2019), мы экспериментировали с несколькими настройками повышения температуры для недостаточно представленных языков, где предложения на языке l, представляющем p_l процентов набора данных, выбираются пропорционально p_l^(1/T)**.** Оптимальная производительность была достигнута при 1/T = 0.3.

Параметры обучения. Наша лучшая модель была обучена с использованием softmax потерь в течение двух эпох с коэффициентом обучения 0.8 и вложениями размерности 256. Мы отбросили слова с менее чем тысячей встреч после повышения температуры и выбрали минимальную и максимальную длину символьных n-грамм в два и пять соответственно, которые были размещены в ячейках размером 1,000,000. Все гиперпараметры были настроены на развивающемся наборе данных Flores-200.

5.1.3 Улучшение LID с помощью лингвистического анализа

Идентификация языка является сложной задачей, где существует множество режимов сбоя, часто усугубляемых разрывом между чистыми данными, на которых обучаются модели LID, и шумными данными, на которых применяются модели LID. Модели LID, обученные в надзорном режиме на грамотно написанных предложениях, могут испытывать трудности при идентификации грамматически неправильных и неполных строк, извлеченных из веба. Кроме того, модели могут легко выучивать случайные корреляции, которые не имеют смысла для самой задачи. В свете этих проблем мы тесно сотрудничали с командой лингвистов на различных этапах разработки LID для выявления правильных направлений работы, смягчения проблем и исследования решений.

Интерфейс проверки LID. Мы использовали линейность fasttext для создания удобного интерфейса, который позволяет лингвистам взглянуть на его внутреннее устройство. Этот инструмент позволяет лингвистам анализировать ошибки модели и выявлять её шаблоны.

Label	Label Score	Text
eng	0.681	(French: Déclaration des droits de l'homme et du citoyen de 1789), set by France's National Constituent Assembly in 1789, is a human civil rights document
fra	0.16	(French: Déclaration des droits de l'homme et du citoyen de 1789), set by France's National Constituent Assembly in 1789, is a human civil rights document

Фигура 8: **Интерфейс проверки LID**, использованный на примере предложения из английской Википедии с коротким отрывком на французском языке. Показаны два наиболее вероятных языка с их оценками. N-граммы, внесшие наибольший вклад (положительный или отрицательный) в прогнозы, выделены (зеленым и красным цветом соответственно).

Как показано на рисунке 8, мы визуализируем вклад каждого n-грамма в окончательное предсказание. В одном из приложений этот инструмент позволил лингвистам заметить сходство в фонотактике между стандартным малайским и индонезийским языками, которые

являются одной из часто путаемых пар языков, и узнать через лингвистические исследования, что несмотря на очевидные различия, существует определенная взаимопонимаемость между ними.

Фильтрация обучающих данных. Для снижения риска обучения модели на спуринговых корреляциях из-за шумных образцов обучения, при моделировании сотен языков мы работали в сотрудничестве с лингвистами над разработкой нескольких фильтров, изображенных в таблице 3 и описанных ниже. Все они последовательно применяются к нашему исходному обучающему набору данных.

• **Фильтрация распределения символов:** Общедоступные наборы данных, которые мы использовали для обучения, в основном были собраны с веб-страниц. При исследовании лингвистами было обнаружено множество случаев ошибочной маркировки предложений, вероятно вызванных наличием коротких отрывков на другом языке на странице, например, индонезийские сайты, на которых публикуются сборники яванских стихов. Мы также заметили случайное творческое использование неожиданных символьных наборов, обычно используемых для украшения или акцентирования внимания, как указано в Caswell et al. (2020). В таблице 3 приведены несколько примеров. Для решения этой проблемы мы искали сдвиги распределения символов, либо вычисляли гистограммы символов, либо рассматривали диапазон Unicode, ожидаемый для языка.

Гистограммы символов: Мы вычисляли распределение символов для каждого языка в нашем наборе разработки и определяли произвольный набор принятых символов для каждого из них, учитывая все символы, попадающие в первый 95-й процентиль. В результате мы фильтровали любое предложение из нашего обучающего набора, состоящее менее чем на 80% из таких принятых символов.

Определение скрипта: Для языков, чей алфавит насчитывает тысячи символов, метод гистограммы символов, упомянутый выше, оказался менее эффективным, поскольку тенденции распределения символов были менее заметными. В качестве альтернативы лингвисты предоставили диапазоны Unicode для определения принятых наборов символов. Любое предложение, содержащее менее 50% символов из этого набора, в конечном итоге отсеивалось. Например, предложения, показанные в таблице 3 для японского и китайского языков, не содержат правильные скрипты Jpan и Hans.

• **Фильтрация для английского языка**: Лингвисты также указали на то, что многие ошибочно помеченные обучающие примеры на самом деле являются простыми английскими предложениями. Это можно объяснить массовым распространением английского языка в сети, даже на страницах, в основном написанных на других языках. Мы создали простой специализированный бинарный классификатор fasttext для фильтрации этих примеров из нашего обучающего набора данных.

Filter	Label	Filtered Sentence
Histogram	urd_Arab	├▇▇▇▇▇═══ℳ😊♓ℳ α Ð α▇ ے یاشاپ ے مضترم :
	dan_Latn	అనంతపూర్ డిస్ట్రిక్ , urdu: عاضہ دابآ لدآ) er et distrikt i den
Script	jpn_Jpan	4.0, CUDA 対応。消費電力は 40W。Quadro FX 380 コア 450MHz
	zho_Hant	容存档于 2009 年 2 月 10 日）． Satellite map 維基衛星
English	tur_Latn	A module is said to be semisimple if it is the sum of simple submodules.
	nld_Latn	Line drawing and design: From the book Brazil and the Brazilians, 1857

Таблица 3: **Примеры предложений, отфильтрованных** из нашего обучающего набора данных для LID.

5.1.4 Результаты

В данном разделе представлено сравнение нашего подхода с существующими общедоступными моделями как на наборе данных Flores-200, так и на аннотированных шумных веб-данных, а также анализ ошибок.

Оценка на наборе данных Flores-200. Мы анализируем производительность наших моделей LID на наборе данных Flores-200 из раздела 4 и сравниваем с другими моделями с открытым исходным кодом. Мы используем Flores-200 для оценки, поскольку текст на целевой стороне проверен людьми и утвержден как написанный на правильном языке. Использование стандартных общедоступных наборов данных для оценки менее надежно, так как они часто содержат ненадежные языковые метки и являются довольно шумными (Kreutzer et al., 2022).

Мы сравниваем нашу модель LID с тремя общедоступными моделями: CLD3, LangId и LangDetect. В таблице 4 представлены показатели производительности нашей окончательной модели LID на множестве различных языковых сочетаний, охваченных всеми четырьмя моделями. В таблице 5 отображены микро F1-оценки и уровни ложных срабатываний по всем языкам, найденным в наборе данных Flores-200. Учитывая различные объемы поддерживаемых языков, мы сообщаем о трех каскадных пересечениях с Flores-200: (1) 51 язык, также поддерживаемых LangId, LangDetect и CLD3, (2) 78 языков, также поддерживаемых LangId и CLD3 и (3) 95 языков, также поддерживаемых CLD3. Мы представляем метрики всех моделей по всем пересечениям, чтобы отразить влияние ложных срабатываний на невидимые языки.

Наша модель способна обрабатывать 200 языков из набора данных Flores-200 (в сравнении с 107 языками, поддерживаемыми CLD3), при этом достигая значительно более высокой производительности по сравнению с LangId, LangDetect и CLD3. Кроме того, повышение микро F1-оценки сопровождается заметным улучшением уровня ложных срабатываний, что указывает на более точное извлечение языков с низкими ресурсами из веб-корпусов (Caswell et al., 2020).

Оценка человеком на шумных веб-данных. Несмотря на сильные результаты на Flores-200, мы ожидаем значительное различие в производительности при применении нашей модели LID к целевым веб-данным.

	# Supported Languages	FLORES-200 ∩ CLD3 ∩ LangId ∩ LangDetect 51 Labels		FLORES-200 ∩ CLD3 ∩ LangId 78 Labels		FLORES-200 ∩ CLD3 95 Labels	
		F1	**FPR**	**F1**	**FPR**	**F1**	**FPR**
LangDetect	55	97.3	0.0526	64.4	0.4503	53.1	0.4881
LangId	97	98.6	0.0200	92.0	0.0874	75.8	0.2196
CLD3	107	98.2	0.0225	97.7	0.0238	97.0	0.0283
Ours	**218**	**99.4**	**0.0084**	**98.8**	**0.0133**	**98.5**	**0.0134**

Таблица 4: **Сравнение моделей идентификации языка с открытым исходным кодом на различных пересечениях меток**. F1 - это микро F1-оценка, а FPR - микро уровень ложных срабатываний.

	Micro F1	Macro F1	Macro Precision	Macro Recall	Macro FPR	Micro FPR
Low-Resource	95.63	95.9	97.6	95.4	0.01213	0.0235
All FLORES-200	95.85	95.5	94.0	95.7	0.02110	0.0210

Таблица 5: **Производительность нашей системы LID на наборе данных FLORES-200**. Лангоиды на арабском и Акан/Тви были объединены после лингвистического анализа.

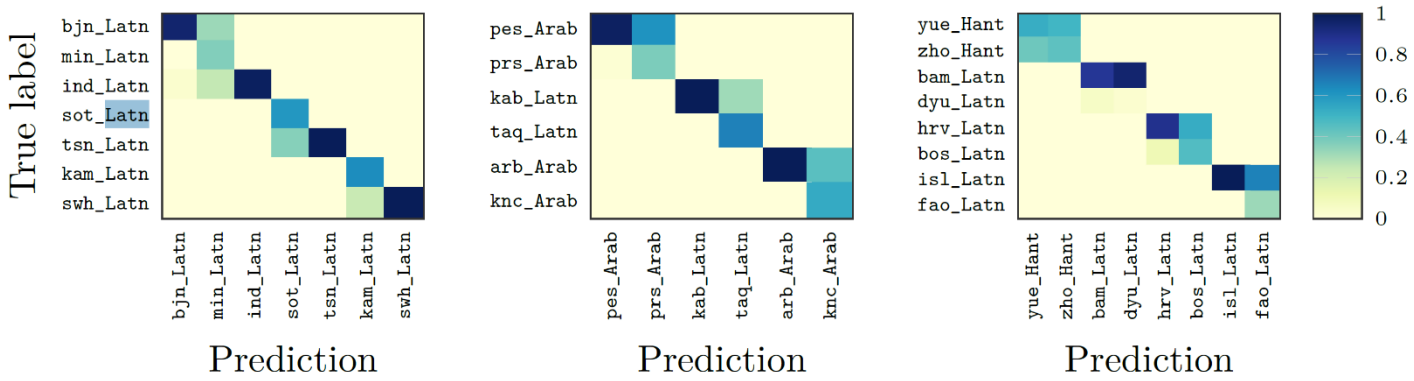
Действительно, различные источники шума, такие как смешение языков, творческое использование различных письменностей и "л33тспик" (использование чисел и символов вместо букв), широко распространены в онлайн-среде. Извлечение предложений с веб-страниц также подвержено неожиданным артефактам, появляющимся после парсинга. Нет готового набора для оценки производительности веб-домена, на основании которого можно должным образом оценить и настроить производительность, не говоря уже о итерациях по выбору конструктивных решений при моделировании. Это побудило нас проверить производительность нашей системы с помощью аннотаторов-людей.

Для этого мы выбрали 74 языка с низкими ресурсами, на которых наш предварительный модуль LID показал низкие значения F1. После первой проверки идентификации языка на веб-данных мы случайным образом выбрали несколько тысяч предложений на различных языках, для которых оценочные баллы предсказания колебались между 50% и 90%. Этот жесткий порог был выбран после ручной проверки, при которой было замечено, что множество ошибок классификации происходило в этом диапазоне. Аннотаторы-люди были задачены с задачей проверки наших случайных предложений и оценки того, действительно ли каждое из них находится в предсказанном языке.

На основе этих аннотаций мы составили набор испытаний для идентификации языка, чтобы проверить нашу окончательную модель LID. Как показано в таблице 6, мы достигли более низкой производительности по сравнению с набором данных FLORES-200, что указывает на значительное несоответствие доменов. Мы также сравниваем производительность с CLD3. Как предложено в работе Caswell et al. (2020), мы сообщаем о неверных срабатываниях (FPR) в дополнение к оценкам F1, чтобы получить лучшее представление о том, как наша модель будет справляться с веб-корпусами с экстремальным дисбалансом классов.

Language	Ours F1	CLD3 F1	Ours FPR	CLD3 FPR
Micro	**79.14**	64.41	**0.79**	1.12
Macro	**74.16**	60.13	**0.77**	1.12

Таблица 6: **Сравнение CLD3 и нашей модели на наборе задач, построенном на основе аннотаций людей**. Показана только средняя производительность языков, поддерживаемых как CLD3, так и нашей моделью. Полная таблица доступна в приложении, Таблица 50.



Фигура 9: **Матрица ошибок на датасете Flores-200** для 21 языка с наименьшей точностью

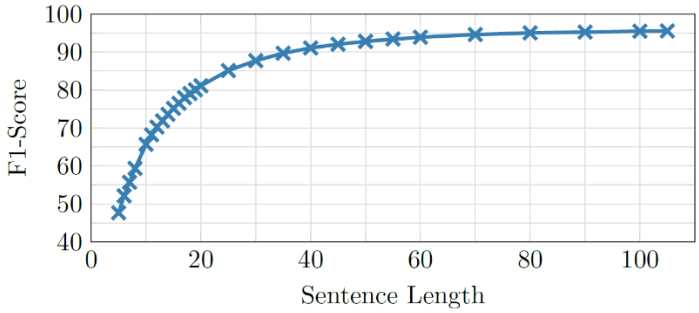
Reference (English)	Asante Twi (aka_Latn) and Akuapem Twi (twi_Latn)
No one was inside the apartment	Na obiara nni odan no mu
There's a tiny cave near the top that must be passed through	ɔbodan ketewa bi bɛn soro ho a ɛsɛ sɛ wofa mu
For a few pennies some children will tell you the story	Sɛ woma mmofra bi sika ketewa bi a, wobɛka anansesɛm yi akyerɛ wo

Таблица 7: **Примеры идентичных предложений на языках** Акан и Тви, двух схожих языках.

Несмотря на расширенное охватом языков, наш подход демонстрирует как более высокий показатель F1 при более низком FPR, что указывает на хорошее соответствие для последующего этапа (раздел 5.2). Мы представляем производительность нашей модели и CLD3 на всех выбранных языках в Таблице 50 в приложении. Следует отметить, что поскольку предложения для аннотации были выбраны на основе предыдущей модели, этот набор вызовов субъективен в связи с использованием промежуточной модели.

Анализ сложных языковых пар. Фигура 9 выявляет небольшую группу пар языков, которые оказались наиболее сложными для нашей системы идентификации языка (LID): Akan/Twi, Dyula/Bambara, Faroese/Icelandic, Western Persian/Dari и Bosnian/Croatian. Мы тесно сотрудничали с лингвистами для их анализа. По результатам проверки мы обнаружили, что эти языковые пары соответствуют высоко схожим языкам с значительным перекрытием в словарном запасе и грамматике. Например, Asante Twi (aka_Latn) и Akuapem Twi (twi_Latn) — два взаимопонятных диалекта языка Akan, которые имеют общие слова, фразы или даже идентичные переводы предложений. Примеры можно найти в Таблице 7. Это свидетельствует о том, что с лингвистической точки зрения путаница в LID, обнаруженная среди этих схожих языков, ожидаема и не является признаком более глубокой проблемы моделирования. На практике это означает, что производительность прогнозирования может быть недооценена для некоторых языков и требует сбора и признания нескольких языковых меток в будущих работах.

Влияние длины предложения. Мы заметили, что прогнозы чаще оказываются более надежными для длинных предложений. Фигура 10 представляет обзор различий в производительности в зависимости от длины входных данных. Это соответствует наблюдению Jauhiainen и др. (2017) и может потребовать дальнейшего изучения. Возможной стратегией смягчения этой проблемы может быть настройка моделей на более сбалансированном разработочном наборе данных по длине. Короткие тестовые предложения могут быть созданы синтетически на основе наших текущих образцов разработки. В нашем текущем подходе мы смягчаем эту проблему путем применения фильтров по длине в монолингвальном конвейере, описанном в разделе 5.2.



Фигура 10: **Влияние длины предложения на производительность**. Мы синтетически создаем тестовые примеры определенной длины без обрезки слов, за исключением языков с непрерывными письменностями.

5.2 Сбор и очистка одноязычных данных в масштабах

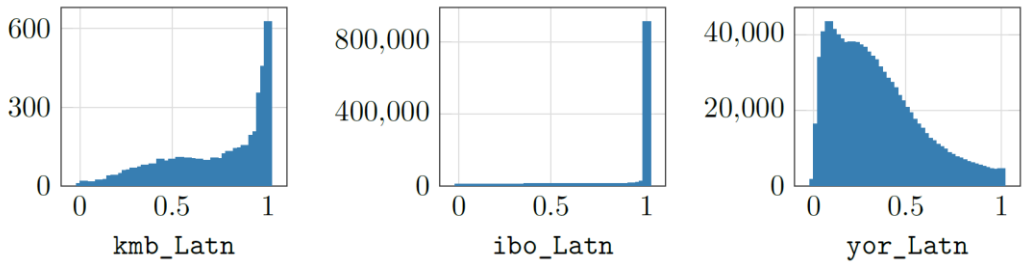
Одноязычные данные являются ценным ресурсом, который можно использовать для различных последующих задач, таких как майнинг битекстов, обратный перевод и обучение языковых моделей. Кроме того, качество данных может существенно повлиять на производительность таких задач. Чтобы максимизировать потенциальные преимущества использования этих источников данных, мы стремимся создавать высококачественные и чистые одноязычные данные. Как обсуждалось ранее в этом разделе, такие данные могут быть дефицитными, особенно в условиях низкоресурсных сред. Поэтому мы решили расширить работы, проведенные в CCNet, CCMatrix (Schwenk et al., 2021b; Wenzek et al., 2020), а также других, таких как OSCAR (Ortiz Suárez et al., 2019). В данном разделе мы описываем наш полный процесс курирования и очистки одноязычных данных.

5.2.1 Описание нашего одноязычного конвейера

Источники данных. Мы начинаем с веб-данных, предоставленных CommonCrawl (CC)18 и ParaCrawl (Bañón et al., 2020). Эти данные были предварительно обработаны для удаления всех разметок и (приблизительной) нормализации кодировки в UTF-8. Чистка HTML преобразует блочные теги в переводы строк, а встроенные теги удаляются. Полученные строки могут содержать много предложений или просто короткий отрывок текста; мы называем их "параграфами".

Применение идентификации языка. Чтобы преобразовать сырой веб-текст в форме параграфов в предложения, мы применяем идентификацию языка иерархически. Сначала мы применяем LID к каждому веб-параграфу. Затем мы используем предсказанный язык для выбора разделителя предложений для этого языка.

Сырые параграфы иногда содержат смесь разных языков или могут включать смену кода. Чтобы избежать смешивания языков, после разделения документов на предложения мы повторно запускаем LID для идентификации языка каждого предложения. Если идентификация языка на уровне предложения не соответствует идентификации на уровне параграфа, мы отбрасываем это предложение, чтобы убедиться, что сохраняем предложения высокой уверенности на целевом языке.



Фигура 11: **Распределение оценок LID на данных ParaCrawl**, иллюстрируемое на примере языков Кимбунду, Игбо и Йоруба.

Мы также отбрасываем предложения, если они не используют ожидаемый алфавит для целевого языка. Следует отметить, что многие предложения содержат значительный уровень шума. В частности, они часто включают длинные URL или списки хэштегов. Это затрудняет процесс идентификации языка и скрипта, так как они на латинском алфавите и не всегда на том же языке, что и исходное предложение. Для определения фактического языка предложения мы усекаем URL и хэштеги перед запуском процесса идентификации языка.

В связи с несоответствием доменов (см. раздел 5.1), наш набор разработки не мог использоваться для настройки пороговых значений нашего классификатора fasttext до желаемого уровня производительности. Вместо этого мы полагались на распределение оценок модели на одноязычных данных набора данных ParaCrawl (Bañón et al., 2020) по всем предсказанным языкам. Мы выбрали этот набор данных, исходя из предположения о том, что его распределение языков реалистично соответствует CommonCrawl, несмотря на неизбежный сдвиг, вызванный использованием модели LID при создании самого ParaCrawl. Большинство языков попали в одно из трех распределений оценок, как показано на рисунке 11 с примерами языков Кимбунду, Игбо и Йоруба.

1. Левоскошенное распределение, где оценки редко опускаются ниже отметки 0.5.
2. Крайне левоскошенное распределение, где почти все оценки выше отметки 0.9.
3. Правоскошенное распределение, где оценки очень редко превышают отметку 0.5.

Этот анализ мотивировал наш выбор порога 0.5 в первых двух случаях как значения по умолчанию, за исключением языков с высокими ресурсами, где мы могли позволить себе более строгие значения, такие как 0.9, и все еще собирать достаточное количество одноязычных данных в нашем последующем конвейере, описанном в разделе 5.2. В последнем случае мы выбирали значения, соответствующие пику каждого распределения (в диапазоне от 0.2 до 0.4), в надежде собрать значительное количество данных из нашего конвейера.

Эвристики для очистки данных. Затем мы применяем несколько эвристик для удаления предложений, которые не соответствуют разумным критериям качества: минимальная и максимальная длина, соотношение пробелов/пунктуации/цифр/эмодзи и максимальное количество повторяющихся символов. Например, если предложение содержит более 20% пунктуации, скорее всего, оно не является хорошо структурированным предложением. Поскольку наши модели-кодировщики (см. раздел 5.3) не были обучены на значительном объеме контента с эмодзи, мы предпочитаем удалить эмодзи из всего текста, чтобы избежать потери предложений, которые могли бы соответствовать, если бы не содержали специальных символов. Некоторые предложения были перенасыщены хэштегами и URL (например, предложения в стиле твитов). В этих случаях мы удаляем эти части предложений и применяем эвристики к усеченному предложению.

Filter	Example	Reason
Low LID threshold	Internet Plus € 58,50	`eng_Latn` at 0.19 LID score
LID mismatch	Best véto ever!	doc. LID French, sent. LID Czech
Numbers	Vol.180 Sep. (2011)	exceeded numbers ratio
Punctuation	. * sApEvAte cHe... » (Previous page)	exceeded punctuation ratio
Emoji	💪💪💪 #gymgirl	exceeded emoji ratio

Таблица 8: **Примеры отфильтрованных предложений** в нашем одноязычном конвейере.

В этих случаях мы удаляем эти части предложений и применяем эвристики к усеченному предложению вместо исходного. Мы отмечаем, что эти соотношения могут различаться в зависимости от языка и не являются универсально применимыми. В некоторых языках концепции могут требовать большего количества слов для передачи, что делает установку порогов на основе длины проблематичной. Аналогично, другие языки могут использовать больше пунктуации или иметь более короткие слова. Поэтому мы не устанавливаем чрезмерно строгие фильтры и изучаем количество отфильтрованного текста для каждого из 200 языков по каждому фильтру. В Таблице 8 приведены типичные примеры отфильтрованных предложений.

Дедупликация. Критерий на основе маржи нашего метода добычи требует уникальных предложений (см. раздел 5.3), но процесс разделения и очистки предложений может порождать множество дублирующегося контента, поэтому мы выполняем глобальный процесс дедупликации для всех предложений на одном языке.

Фильтрация языковой моделью. Поскольку нас интересует сохранение высококачественных предложений в наших наборах данных для последующего обучения наших окончательных многоязычных моделей перевода, когда это возможно, мы также применяем фильтрацию с использованием языковой модели (LM). На практике трудно обучать высококачественные языковые модели для языков с низкими ресурсами, поэтому мы сосредотачиваемся на применении фильтров языковых моделей только для нескольких языков с высокими ресурсами. Поскольку большинство наших данных добывается с одной стороны пары с использованием английского языка, мы считаем, что если у нас есть высококачественный контент в корпусе на английском языке, процесс выравнивания добычи также выдаст высококачественный контент на другой стороне пары. Для английского языка мы используем модель KenLM (Heafield, 2011) из CCNet (Wenzek et al., 2020).

Вычислительные вызовы. Мы обработали около 37,7 петабайт данных через весь конвейер. Это было вызовом для управления данными и использования дискового пространства. В частности, нам пришлось принимать сложные решения при фильтрации языков с высокими ресурсами, чтобы искусственно сохранить только около 30% данных из наиболее объемных языков. Этот порог был определен как предел, ниже которого оценка LM идентифицирует предложения низкого качества. Обработка была распределена по множеству машин и заняла несколько месяцев, чтобы получить окончательный одноязычный набор данных. Мы понимаем, что обработка такого объема данных не всегда возможна, и открыто предоставляем значительную часть наших результатов и кода для удобства использования всеми желающими.

5.2.2 Одноязычные данные в кратком обзоре

Мы начали с 107,9 миллиарда абзацев из интернета (97,4% высокоресурсные, 2,6% низкоресурсные) и отбросили 2,8 миллиарда предложений (85,3% высокоресурсные, 14,7% низкоресурсные) из-за несоответствия LID/скрипта или низкого значения LID (см. раздел 5.1), чтобы получить 43,7 миллиарда одноязычных предложений (90,8% высокоресурсные, 9,2% низкоресурсные) для использования в добыче. Из них 21,5 миллиарда предложений на английском языке. Для подробностей по объему данных см. Таблицу 9. Обратите внимание, что такое значительное снижение числа предложений связано с тем, что мы значительно фильтровали языки с высокими ресурсами, чтобы сохранить топ 30% предложений на основе оценки языковой модели (LM score).

	Total	Min		Max		Median		Mean
Low-Resource								
Raw Data (Para.)	2.4B	27.1K	tzm_Tfng	465.8M	nob_Latn	3.3M	mai_Deva	17.9M
LID/Script Mismatch (Sent.)	0.3B	0.2K	tzm_Tfng	47.8M	nob_Latn	0.3M	fao_Latn	2.2M
Clean Sentences	3.6B	1.3K	tuk_Latn	330.3M	glg_Latn	4.4M	tso_Latn	26.8M
High-Resource								
Raw Data (Para.)	105.4B	4247.8K	tsn_Latn	73.2B	eng_Latn	83.1M	eus_Latn	3401.2M
LID/Script Mismatch (Sent.)	2.5B	124.4K	ben_Beng	1.6B	eng_Latn	5.0M	als_Latn	81.2M
Clean Sentences	40.1B	5153.8K	tsn_Latn	21.5B	eng_Latn	234.3M	als_Latn	1294.5M

Таблица 9: **Основные статистические данные** по обработке одноязычных данных с низким и высоким ресурсом.

5.3 Добыча битекстов для языков с низкими ресурсами

Машинный перевод, как и многие другие приложения машинного обучения, сильно зависит от данных. Предыдущие работы четко установили, что качество перевода обычно повышается с увеличением количества высококачественных обучающих данных (Koehn and Knowles, 2017). Существующие параллельные корпуса для языков с низкими ресурсами часто опорочиваются на известных коллекциях многоязычных контентов, таких как Библия или публикации международных организаций. Они часто ограничены по объему и области применения. В данном разделе мы описываем, как автоматически создаем наборы данных для обучения перевода для языков с низкими ресурсами через добычу битекстов. Мы в основном сосредотачиваемся на битекстах, сопоставленных с английским, но также интересуемся добычей через другие языковые пары, так как было показано, что это может улучшить общую производительность многоязычной системы перевода (Fan et al., 2020).

Что такое добыча битекстов? Основная идея нашего подхода к добыче битекстов заключается в первоначальном изучении многоязычного пространства вложений предложений и использовании меры сходства в этом пространстве для определения, параллельны ли два предложения или нет. Это сравнение можно выполнить для всех возможных пар в двух коллекциях одноязычных текстов, называемых глобальной добычей в работе Schwenk et al. (2021a). Другой распространенный альтернативный подход называется иерархической или локальной добычей и состоит в том, чтобы сначала выполнить выбор потенциальных пар документов, а затем ограничить добычу предложениями внутри каждой пары документов. Проект European ParaCrawl является типичным примером такого подхода (Bañón et al., 2020). Ранее проводившиеся работы по этим направлениям занимались как веб-сайтами с параллельными текстами (Resnik and Smith, 2003), так и сопоставимыми данными (Fung and Cheung, 2004). В данной работе мы следуем подходу, использованному в WikiMatrix (Schwenk et al., 2021a) и CCMatrix (Schwenk et al., 2021b), которые оба используют глобальную добычу.

Так как наш подход к добыче требует многоязычного пространства вложений предложений, мы столкнулись с несколькими вызовами при масштабировании этого представления до 200 языков инициативы No Language Left Behind. Например, как мы можем убедиться, что все языки хорошо изучены? И как следует учитывать большие дисбалансы доступных обучающих данных? Обучение массово многоязычного кодировщика предложений с нуля каждый раз при добавлении нового набора языков было бы вычислительно очень дорогостоящим. Кроме того, такой подход имеет недостаток в том, что изученные пространства вложений от каждой новой модели не (естественно) совместимы между собой. Это может сделать добычу нецелесообразной, так как для каждого нового кодировщика

потребуется перевложение всего доступного одноязычного материала, и только для английского это означает десятки миллиардов предложений и большие вычислительные ресурсы. Для преодоления этих проблем одним из подходов является обучение меньших взаимно совместимых кодировщиков предложений с использованием техники дистилляции учителя-студента, предложенной Reimers and Gurevych (2020). Несколько расширений этой основной идеи были предложены Heffernan et al. (2022), которые мы приняли на вооружение (см. раздел 5.3.2).

5.3.1 Связанные работы

Методы добычи. Для поиска сопоставленных текстов ранние подходы фокусировались на информации, выходящей за пределы самих текстов. Одним из заметных примеров является алгоритм STRAND, который искал статьи с аналогичной структурой документа для нахождения переведенных веб-страниц (Resnik, 1999; Resnik and Smith, 2003). В дальнейшем работы использовали тексты на страницах в качестве основы для выравнивания с использованием таких техник, как кросс-языковый поиск документов (Munteanu and Marcu, 2005; Utiyama and Isahara, 2003), мешок слов или языковые модели (Buck and Koehn, 2016), сходство Жаккара (Azpeitia et al., 2017, 2018; Etchegoyhen and Azpeitia, 2016) или перевод (Abdul-Rauf and Schwenk, 2009; Bouamor and Sajjad, 2018). Более современные подходы начали использовать преимущества обучения представлений, кодируя тексты в пространство вложений и затем используя методы на основе расстояний для определения, имеют ли пара текстов на разных языках схожее значение. Работы, такие как España-Bonet et al. (2017); Guo et al. (2018); Hassan et al. (2018); Yang et al. (2019), использовали двуязычные вложения, но у них было ограничение на прямую добычу данных через множество языков. Для решения этой проблемы обучение массово многоязычного пространства вложений позволяет кодировать и добывать данные для любой пары языков (Artetxe and Schwenk, 2019a,b; Feng et al., 2020; Kvapilíková et al., 2020; Schwenk, 2018). При попытках масштабной добычи были применены несколько подходов к большому количеству языковых пар. Один из таких примеров - проект ParaCrawl20, который добывал данные для всех официальных языков ЕС. Аналогичный подход представлен в проекте ccAligned (El-Kishky et al., 2020). Недавно Schwenk et al. (2021b) представили CCMatrix, который успешно добывал миллиарды предложений из веба, используя многоязычное вложение LASER (Artetxe and Schwenk, 2019b). Проект Samanantar сосредоточился на предоставлении большого добытого корпуса для одиннадцати индийских языков (Ramesh et al., 2022).

Обучение представлений предложений на многих языках. Существует широкий спектр работ, посвященных обучению мультиязычных представлений, таких как mBERT (Devlin et al., 2019), XLM (Conneau and Lample, 2019) и XLM-R (Conneau et al., 2020). Однако при применении этих подходов для получения представлений на уровне предложения часто возникает проблема отсутствия явного критерия на уровне предложения во время обучения, что ведет к низкой производительности на задачах, таких как поиск битекстов (Hu et al., 2020). Для решения этой проблемы были исследованы методы с явной целью на уровне предложения, такие как SentenceBERT (SBERT) (Reimers and Gurevych, 2019). SBERT основан на сети типа "сиамских близнецов", которая впоследствии дообучается на данных о логической несоответственности (NLI) (Bowman et al., 2015) и создает вложения предложений путем классификации пар предложений. Аналогично этой сети типа "сиамских близнецов", LaBSE также использует подход с двойным кодированием с использованием BERT.

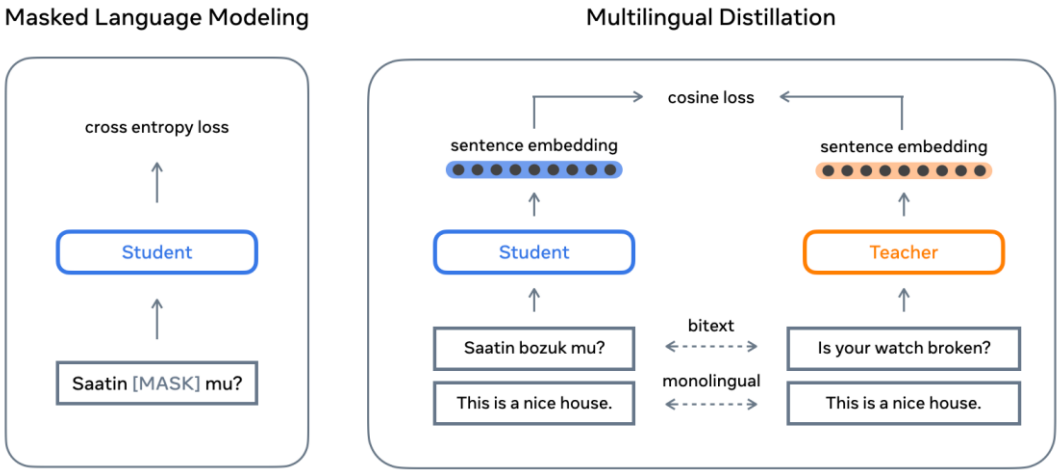


Рисунок 12: **Архитектура метода учителя-студента LASER3**. Для получения более подробной информации обращайтесь к работе Heffernan et al. (2022).

Однако предварительное обучение для кодировщиков BERT выполняется с использованием как задачи маскированного моделирования языка (MLM), так и задачи моделирования перевода (TLM) (Conneau et al., 2020). Предложение встраивания предложений осуществляется путем прохождения двухязычных пар перевода через двойную модель кодировщика, а затем применения потерь аддитивного мягкого маржи (Yang et al., 2019). Еще одна популярная модель мультиязычных вложений - LASER (Artetxe и Schwenk, 2019b). В отличие от предыдущих подходов на основе BERT, здесь нет предварительного обучения, и встраивание предложений осуществляется с использованием архитектуры кодировщик-декодировщик, а затем максимальное объединение по выводам кодировщика.

При попытках изучения мультиязычного пространства встраивания одним из ограничений многих существующих подходов является то, что каждый раз, когда модель должна быть расширена для включения нового набора языков, вся модель должна быть переобучена с нуля, что очень дорого. Для решения этой проблемы Wang et al. (2020b) предложили технику, которая позволяет расширять mBERT на языки с недостаточными ресурсами, сначала увеличивая размер существующего словаря, а затем продолжая обучение без учителя с использованием монолингвистических данных с низкими ресурсами. Другой подход, представленный Reimers и Gurevych (2020), использует подход учителя-ученика. В этой схеме дистилляции используются англоязычные битексты как для изучения английского пространства встраивания учителя (SBERT), так и для изучения нового языка на стороне не-английской. Heffernan et al. (2022) далее

разработали этот подход, экспериментируя с различными архитектурами для учителя и ученика (например, BiLSTM и Transformer), а также применяя такие дистиллированные представления предложений к задаче майнинга битекстов.

5.3.2 Подход учителя-студента к майнингу

Общий подход фокусируется на начале с учителем модели кодировщика массово мультиязычных предложений и ее адаптации к нескольким различным моделям студентов с низкими ресурсами. Это позволяет нам добавлять языки с недостаточными ресурсами без необходимости конкурировать с языками с высокими ресурсами за мощность, избегать переобучения полной модели с нуля и поддерживать совместимость мультиязычных вложенных пространств для последующего майнинга. Рисунок 12 подводит итоги общей архитектуры подхода учителя-студента. Учитель, LASER2, является улучшенной версией открытого кодировщика LASER. Оригинальная процедура обучения (Artetxe и Schwenk, 2019b) была изменена следующим образом: использование токенизации SentencePiece и увеличение образцов для языков с низкими ресурсами. Архитектура 5-слойного BiLSTM кодировщика и метод максимального объединения для получения встраиваний предложений оставались неизменными. Обучение проводилось на том же наборе из 93 языков с общедоступными ресурсами, полученными из OPUS (Tiedemann, 2012). Для более подробной информации по оригинальной процедуре обучения LASER читателю рекомендуется обратиться к работе Artetxe и Schwenk (2019b). Обучение студентов следует описанному подходу более подробно в работе Heffernan et al. (2022):

- студенты специализируются на одном языке или нескольких схожих языках;
- студенты инициализируются случайным образом, поскольку мы работаем с языками с недостаточными ресурсами, для которых нет предварительно обученной языковой модели;
- у студентов может быть специальный словарь SentencePiece, отличающийся от учителя, чтобы лучше адаптироваться к скриптам и токенам в языках студентов (см. раздел 5.3.3);
- студенты учатся минимизировать косинусное расстояние с учителем, так как мы также используем косинусное расстояние для майнинга битекстов (см. Рисунок 12);
- у студентов может быть потеря MLM для использования монолингвистических данных студентов (см. Рисунок 12 и раздел 5.3.3).

Параметры обучения. Наши студенческие энкодеры использовали 12-слойный трансформер с размером скрытого слоя 1024 и 4 внимательных головы, всего около 250 миллионов параметров. Все студенты обучались с использованием доступных битекстов на их родном языке, дополненных двумя миллионами предложений на английском и испанском языках. Мотивация заключается в якорении студентов в английском пространстве вложений, укреплении его за счет включения англо-испанских битекстов из CCMatrix и совместном изучении новых языков. Эта техника особенно полезна, когда доступны ограниченные объемы битекстов для обучения студентов. Обучение учителя-студента проводилось на 16 GPU с оптимизатором ADAM, скоростью обучения 0.0005 и размером пакета 10,000. Мы обучили студенческие энкодеры для 148 языков и назвали эти модели LASER3.

Прокси-метрика для новых энкодеров. Добытые битексты будут последующим образом использованы для улучшения качества машинного перевода для 200 языков. Следовательно, нашей основной метрикой является качество нейронного машинного перевода (NMT). Однако добыча данных и обучение NMT требуют значительных вычислительных ресурсов, и систематическое выполнение этой оценки для множества вариантов энкодеров представляется невозможным. В качестве прокси-оценки мы используем ошибку мультиязычного поиска схожести на основе добычи, называемую здесь xsim. В отличие от косинусной точности, которая выравнивает вложения на основе наивысшего косинусного значения, xsim выравнивает исходные и целевые вложения на основе наивысшего оценочного значения, что было показано полезным в добыче (Artetxe и Schwenk, 2019a). Оценка на основе маржи определяется следующим образом:

$$\text{score}(x, y) = \text{margin}\left(cos(x, y), \sum_{z \in NN_k(x)} \frac{cos(x, z)}{2k} + \sum_{v \in NN_k(y)} \frac{cos(y, v)}{2k}\right) \qquad (1)$$

где x и y - это исходные и целевые предложения, а NNk(x) обозначает k ближайших соседей x на другом языке. Мы устанавливаем k равным 4. Все результаты xsim вычисляются на наборе данных Flores-200 devtest, используя отношение маржи, где margin(a; b) = a / b. Кроме того, все оценки рассчитаны относительно английского языка (т.е. xxx ! eng). Английский язык кодируется учителем, а другой язык - студентом LASER3. Для упрощения дальнейших исследований с использованием xsim, мы также открываем этот метод оценки для сообщества.22

Оценка энкодеров end-to-end. Как только мы определили лучший энкодер предложений для каждого языка, используя оценки xsim, мы проводим добычу, добавляем полученные данные к существующим битекстам и обучаем двуязычную систему машинного перевода (NMT). Первичные эксперименты показали, что порог на марже 1.06 является лучшим компромиссом между точностью и полнотой для большинства языков. Для этих базовых значений NMT мы не применяем дополнительной фильтрации битекстов и оставляем это на этап обучения нашей массово-мультиязычной системы NMT (см. раздел 6.4). Таким образом, качество перевода сравнивается с обучением той же системы NMT только на существующих битекстах, что позволяет оценить улучшения, достигнутые только благодаря добытым битекстам. Мы дополнительно ограничиваем наши эксперименты переводом с иностранных языков на английский. Это помогает сравнить ресурсы и оценить качество перевода между всеми языками. В английском корпусе, который мы добываем, содержится 21.5 миллиарда уникальных предложений.

Мы не пытаемся оптимизировать архитектуру и параметры двуязычных систем NMT под характеристики каждой пары языков, такие как размер доступных битекстов, а используем одну и ту же архитектуру для всех. Поэтому сообщаемые результаты следует рассматривать не как лучшие возможные в условиях доступных ресурсов - они представляются в основном для валидации добытых битекстов. Мы используем 12-слойный кодировщик и декодировщик и обучаем на 100 эпохах. Словарь SentencePiece (SPM) составляет 7,000 токенов. Мы ищем лучшую производительность на развивающемся наборе данных Flores-200 и сообщаем о детокенизированном BLEU на Flores-200 devtest.

5.3.3 Обучение языково-специфических энкодеров

Оригинальный энкодер LASER, использованный в проекте CCMatrix (Schwenk et al., 2021b), хорошо справляется с несколькими языками с высокими ресурсами, такими как арабский, китайский, чешский, немецкий, японский, польский или индонезийский. Мы использовали битексты CCMatrix напрямую для этих языков. Затем мы обучили новые энкодеры предложений и провели добычу для 148 оставшихся языков. В следующем разделе мы обсудим несколько представительных индивидуальных языков и семейств. Для проведения добычи мы применили те же алгоритмы и оптимизации, что и предложено в работе Schwenk et al. (2021b).

Улучшение LASER. Для подчеркивания улучшений LASER с использованием нашего подхода учителя и ученика, на рисунке 13 показаны ошибки xsim для некоторых поддерживаемых LASER языков, которые были переобучены с использованием наших моделей ученика LASER3. В среднем ошибка xsim снизилась с 61 до 0.91. В частности, некоторые языки, такие как бирманский (mya_Mymr) и ирландский (gle_Latn), показали самые значительные снижения xsim, с падением с 93.3 до 0.89 и с 92.5 до 0.79, соответственно.

Европейские меньшинственные языки. В Европе существует большое разнообразие языков, которые говорятся местно в различных регионах. Здесь рассматриваются 10 языков. Для большинства из них мы не смогли найти значительные объемы общедоступных битекстов и сильно полагаемся на NLLB-Seed.

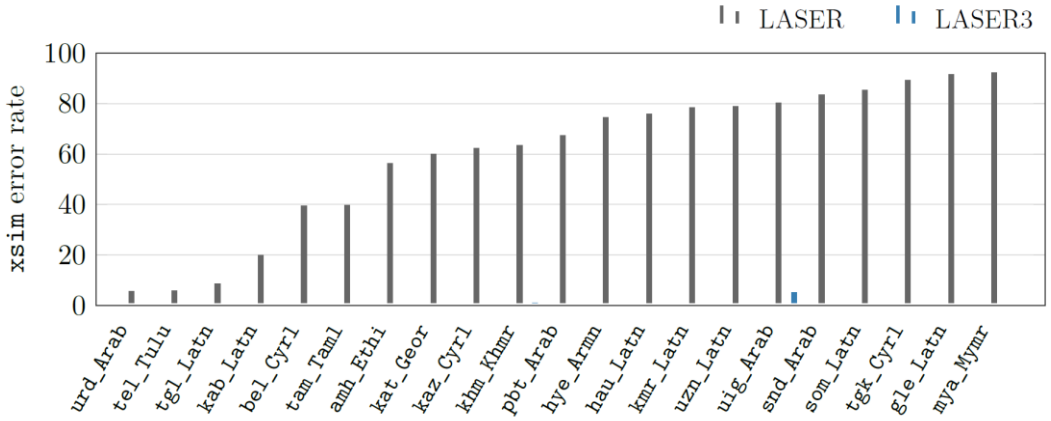


Рисунок 13: **Сравнение уровней ошибок xsim** для LASER3 и оригинального кодировщика LASER. LASER3 показывает значительные улучшения (меньшие значения лучше).

Lang.	fao	fur	lij	lim	lmo	ltz	srd	szl	vec	ydd
Script	Latn	Latn	Latn	Latn	Latn	Latn	Latn	Latn	Latn	Hebr
Addtl. Lang	deu	ita	ita	nld	ita	deu	ita	pol	ita	deu
Bitexts [k]	6.6	6.3	2.2	5.4	1.3	9.8	1.4	6.4	1.2	6.2
BLEU	0	0	0	0	0	0	0	0	0	0
xsim [%]	2.57	0.1	0.2	16.1	1.09	0.59	0.1	0.69	2.77	0.1
Monolingual	1.2M	737k	106k	15M	61M	123M	515k	2.5M	12M	12M
Mined	1.6M	532k	631k	2.0M	4.1M	5.5M	723k	1.0M	2.5M	3.3M
BLEU	10.6	23.5	13.4	5.5	20.7	37.0	20.9	18.9	17.8	30.1

Таблица 10: **Статистика и производительность майнинга** для европейских языков меньшинств. Оценки BLEU представляют перевод с иностранных языков на английский.

(см. Раздел 4.2). Поэтому мы выбрали обучать индивидуальные кодировщики предложений, специфичные для каждого из этих языков, и дополнили тренировочные данные 2 миллионами битекстов на аналогичных высокоресурсных европейских языках. Типичными примерами являются сицилийский в паре с итальянским или силезский в паре с польским. Подробная информация по всем 10 языкам приведена в Таблице 10. Мы наблюдаем, что объединение языка меньшинства с другим аналогичным крупным языком дает кодировщики с очень низкими уровнями ошибок xsim для большинства языков, и мы смогли найти большое количество битекстов, что привело к хорошей производительности НМТ. Три языка достигают оценки BLEU, превышающей 20 (fur_Latn, lmo_Latn и srd_Latn), и два языка превышают 30 (ltz_Latn и ydd_Hebr).

Креольские языки. Мы применили аналогичную стратегию для креольских языков. С лингвистической точки зрения, креольские языки могут быть результатом контактной ситуации между языками на протяжении довольно короткого времени (Lent и др., 2022). При обучении кодировщиков предложений мы сочетали каждый креольский язык с «похожим» высокоресурсным языком (см. Таблицу 11). Кодировщики предложений имеют очень низкий уровень ошибок xsim для всех языков, за исключением санго (sag_Latn), для которого мы не смогли найти аналогичный язык с достаточными ресурсами.

Lang.	hat	kea	pap	sag	tpi
Script	Latn	Latn	Latn	Latn	Latn
Addtl. Lang	fra	por	spa por	lin	eng
Bitexts	334	6	5	282	458
BLEU	20.2	0	0	4.8	14.7
xsim [%]	1.19	1.19	0.1	8.6	0.2
Monolingual	14M	227k	28M	645k	1.7M
Mined	8.0M	656k	7.3M	1.9M	1.2M
BLEU	29.2	4.9	40.9	5.3	16.1

Таблица 11: **Статистика и производительность майнинга** для креольских языков.

Lang.	kab	taq	taq	tzm
Script	Latn	Latn	Tfng	Tfng
bitexts	72k	10.2k	4k	6.2k
BLEU	1.2	0	0	0
xsim [%]	0.99	24.11	35.57	3.66
Monolingual	3.4M	23k	5k	59k
Mined	3.1M	240k	-	111k
BLEU	6.2	1.2	-	3.8

Таблица 12: **Статистика и производительность майнинга** для берберских языков.

Давайте сравним два крайних случая: папьяменто (pap_Latn) и кабовердиану (kea_Latn). Для первого мы смогли собрать 28 миллионов предложений одноязычных данных, в то время как для второго у нас меньше 300 тысяч. Оба языка имеют менее десяти тысяч предложений существующих битекстов, что явно недостаточно для обучения системы НМТ. Мы смогли найти более 7 миллионов битекстов для папьяменто, что дало впечатляющий результат BLEU 40.9, в то время как для кабовердиану мы достигли лишь BLEU 4.9. Это подчеркивает, что количество доступных одноязычных данных критично для успешного майнинга битекстов.

Берберские языки. Мы рассмотрели несколько языков берберской семьи, а именно кабильский (kab_Latn), тамашек (taq) и центральный атласский тамазигхт (tzm). Мы рассматриваем тамазигхт, написанный на письме тифинаг, и тамашек как на латинском, так и на тифинаг. Все эти языки имеют очень ограниченные ресурсы, с едва ли 10,000 предложений доступных битекстов и 72,000 для кабильского. Также было очень сложно собрать одноязычные данные для этих берберских языков. Эти статистические данные приведены в Таблице 12.

Мы смогли найти битексты для кабильского языка и достигли скромного результата BLEU 6.2. Уровни ошибок xsim для тамашека превышают 20%, и недостаточное количество одноязычных данных делает невозможным найти битексты хорошего качества для тамашека. Для тамазигхта удалось найти очень небольшое количество битекстов. Тамашек и тамазигхт являются типичными примерами языков с очень ограниченными ресурсами, для которых, кажется, очень трудно собрать письменные материалы для поддержки обучения системы машинного перевода.

Малайско-полинезийские языки. Рассмотрим более крупную семью языков: малайско-полинезийскую семью. Мы обсуждаем здесь 13 языков из этой семьи и используем один кодировщик для десяти из них. Языки фиджийский, маори и самоанский обрабатываются отдельным кодировщиком. Обзор результатов приведен в Таблице 13.

Мы наблюдаем очень низкие уровни ошибок xsim для большинства языков, хотя несколько языков имеют менее ста тысяч предложений битекстов. Обучение всех языков вместе с использованием одного специфического кодировщика для этой семьи, кажется, очень полезно для этих языков с очень ограниченными ресурсами. Кроме того, нам удалось собрать несколько миллионов предложений одноязычного текста для большинства языков. Это создает оптимальные условия для майнинга, и мы достигаем значительных улучшений в оценке BLEU по сравнению с обучением двуязычной системы НМТ только на публичных битекстах.

Мы имеем два языка, которые улучшили свои показатели более чем на двадцать пунктов BLEU: яванский (jav_Latn): с 11.1 до 31.2 и суданский (sun_Latn): с 1.5 до 28.5. Даже языки с большими ресурсами, такие как тагальский (tgl_Latn) и стандартный малайский (zsm_Latn), показывают значительные улучшения в оценке BLEU. Мы не добывали новые битексты для индонезийского, так как CCMatrix уже предоставляет 70 миллионов параллельных предложений, но мы включили его для помощи в обучении кодировщика предложений для этих языков. Мы также наблюдаем очень хорошие результаты для маори: оценка BLEU улучшилась с 5.6 до 20.5, в то время как добытые битексты не улучшили производительность НМТ для фиджийского.

Lang.	bitexts	BLEU	xsim %	Monol.	Mined	BLEU
ace_Latn	39.2k	0	2.37	2.2M	1.4M	10.3
bug_Latn	21.8k	0	1.58	0.7M	717k	4.2
ceb_Latn	1.1M	34.4	0.1	23.6M	8.1M	39.0
ind_Latn	11M	-	0.1	-	-	-
jav_Latn	86k	11.1	0.1	27.2M	8.5M	31.2
pag_Latn	327k	15.6	0.69	3.9M	1.9M	18.5
sun_Latn	32.3k	1.5	0.59	8.2M	6.1M	28.5
tgl_Latn	1.3M	40.2	0.1	89M	33M	43.8
war_Latn	331k	26.5	0.2	26.9M	4.9M	36.5
zsm_Latn	2.3M	34.4	0.0	640M	40.5M	41.4
fij_Latn	667k	15.0	0.3	1.5M	2.2M	14.6
mri_Latn	45k	5.6	1.38	3.8M	3.2M	20.5
smo_Latn	419k	20.0	0.2	3.5M	3.7M	25.6

Таблица 13: **Статистика и производительность майнинга** для малайско-полинезийских языков.

Африканские языки. Среди 200 языков проекта NLLB, 55 являются языками, на которых говорят на африканском континенте, что составляет более четверти всех языков, которыми мы занимаемся. За исключением семи языков — современного стандартного арабского, африкаанс, южного сото, суахили, тсвана, коса и зулу — все они являются языками с ограниченными ресурсами, то есть с менее чем одним миллионом доступных пар предложений. Двенадцать из них даже имеют менее ста тысяч пар предложений, которые мы назвали языками с очень ограниченными ресурсами. Кроме того, нам было трудно собрать значительные объемы одноязычных данных. Учитывая эти факты, обучение кодировщиков предложений для африканских языков и майнинг качественных битекстов

оказалось серьезной проблемой — даже для более широкой научной среды. Фактически, сбор ресурсов, обучение систем НМТ и проведение оценок для африканских языков является фокусом нескольких работ (Abbott and Martinus, 2019; Azunre et al., 2021c; Dabre and Sukhoo, 2022; Emezue and Dossou, 2020; Hacheme, 2021; Nekoto et al., 2020; Siminyu et al., 2021). Подробное описание и анализ наших усилий представлено в работе Heffernan et al. (2022).

Средняя оценка BLEU для 44 языков увеличилась с 11.0 до 14.8 благодаря добытым битекстам. Мы также внедрили новую процедуру обучения, которая сочетает в себе контролируемое обучение, то есть минимизацию косинусной потери между учителем и учеником, и неконтролируемое обучение с маскировкой LM (см. левую часть Рисунка 12). Это позволило нам воспользоваться одноязычными данными во время обучения кодировщика. Этот новый подход дал улучшенные кодировщики для сложных языков, таких как волоф (Heffernan et al., 2022).

Training	SPM	#train	amh_Ethi	tir_Ethi
LASER2	50k joint	220M	34.9	92.9
Semitic	50k joint	9M	0.2	1.19
Ge'ez	8k specific	0.7M	0.1	0.89
LaBSE	501k joint	≈ 6000M	0	13.74

Таблица 14: **Уровни ошибок xsim на тестовом наборе данных FLORES для амхарского и тигринья, а также различные стратегии обучения.** Указанное количество обучающих данных исключает 4 миллиона предложений на английском языке для наших моделей.

Обработка специфических письменностей. В массово-многоязычных системах распространенным подходом является использование одного и того же словаря SentencePiece. В наших первоначальных экспериментах мы повторно использовали словарь на 50 тыс. токенов из LASER2 для всех кодировщиков учеников. Это имеет некоторые преимущества, так как общий словарь может облегчить обобщение, когда одни и те же токены встречаются в нескольких языках. Однако, несмотря на увеличение выборки данных, языки с ограниченными ресурсами могут плохо представляться в общем словаре. Затем мы исследовали использование специфических словарей для небольших подмножеств языков. В Таблице 14 суммированы результаты для различных стратегий словарей.

Амхарский язык был частью LASER2, но уровень ошибок xsim достаточно высокий, и LASER2 плохо обобщается на тигринья. Сначала мы исследовали обучение кодировщика для трех семитских языков: амхарский, тигринья и мальтийский. Это привело к значительному улучшению: xsim=0.2% и 1.19% соответственно, что подчеркивает полезность обучения учителя-ученика и специфических кодировщиков для небольшого набора схожих языков. Затем мы обучили кодировщик только для амхарского и тигринья, связав их с английским, как и во всех наших экспериментах, с использованием специфического словаря SentencePiece на 8 тыс. токенов для лучшей поддержки геэзского алфавита. Это снизило уровень ошибок xsim до 0.1% и 0.89% соответственно, даже при использовании меньшего объема обучающих данных. Наша лучшая модель на уровне с LaBSE (которая включает только амхарский), и значительно превосходит ее для тигринья.

5.3.4 Майнинг в кратком изложении

В общей сложности, мы добыли 148 битекстов, связанных с английским языком, что составляет 761 миллион пар предложений с оценочным коэффициентом выравнивания не менее 1.06. Как упоминалось ранее, битексты, выровненные с английским для остальных языков, были взяты из CCMatrix (Schwenk et al., 2021b). Мы также предоставляем 1465 пар неанглоязычных битекстов. Это соответствует 302 миллионам пар предложений. Это включает все языковые пары среди семей африканских, индийских и малайско-полинезийских языков, а также выравнивания всех африканских языков с французским.

5.3.5 Ограничения крупномасштабного майнинга

Отмечается, что для некоторых языков мы можем создать только небольшое количество битекста через майнинг данных. Главным ограничивающим фактором является в основном отсутствие одноязычных данных. Многие языки с ограниченными ресурсами имеют ограниченное присутствие в сети, и данные, которые мы курируем, могут быть сильно фильтрованы на многих этапах: идентификация языка, агрессивная очистка одноязычных данных или даже плохо выровненный битекст. Даже с нашими лучшими усилиями эти вызовы усиливаются, и они могут сильнее повлиять на определенные языки по сравнению с другими. Кроме того, следует отметить, что даже в наших добытых битекстах мы все еще включаем некоторое содержание, которое уже доступно из других источников — хорошим примером является Христианская Библия. Важным финальным аспектом является то, что в сети уже есть машинные переводы контента. Например, многие веб-сайты могут использовать перевод для «интернационализации» своего контента. Положительным аспектом является то, что большинство языков, на которых мы сосредоточены, отсутствуют в большинстве существующих коммерческих сервисов перевода (см. Таблицу 1), однако при майнинге против языков с более высокими ресурсами вероятно, что наши добытые наборы данных содержат переведенный контент.

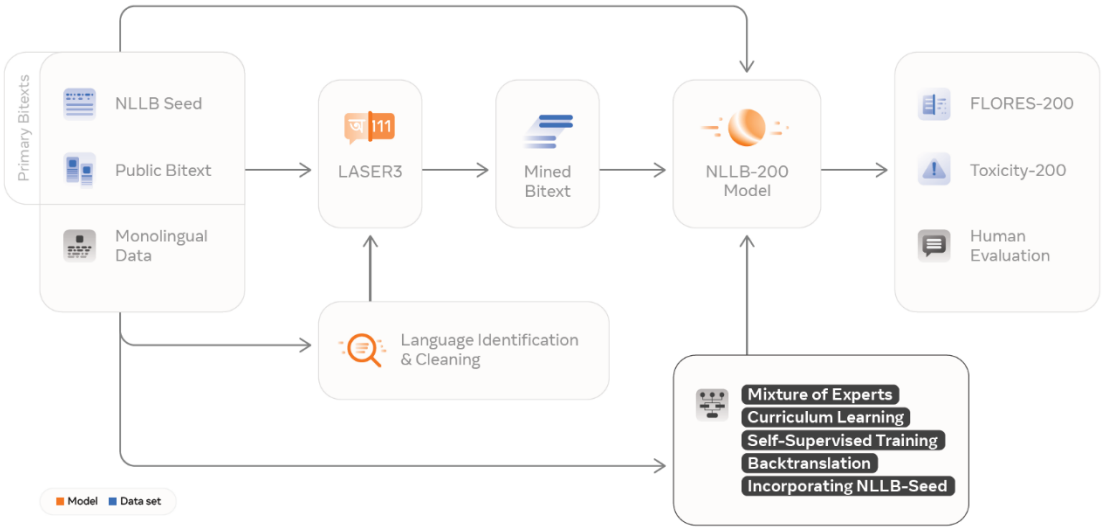
5.3.6 Этические аспекты исследования майнинга

Для завершения данного раздела мы хотим обратить внимание на проблему владения данными. Несмотря на то, что мы провели необходимую работу по развертыванию всех лицензированных наборов данных, вопрос владения данными низкоресурсных языков остается открытым. В нашем исследовании многие заинтересованные стороны с низкими ресурсами выражают мнение, что обмен доступом к языковым данным может, на самом деле, быть необходимым компромиссом для технологического развития. Блокировка такого доступа означает блокировку любых будущих выгод, которые могли бы положительно сказаться на сообществах с низкими ресурсами языков. Однако мы подчеркиваем, что доступ и владение — это два различных концепта. Несмотря на то что мы используем множество наборов данных для языков с ограниченными ресурсами, владение в конечном итоге принадлежит носителям этих языков.

5.4 Заключение

Мы описали нашу методологию для автоматического создания выровненных тренировочных данных для языков с ограниченными ресурсами. Мы сталкиваемся с значительными вызовами, так как битекстовые данные для обучения кодировщиков предложений, таких как LASER, и одноязычные данные для майнинга являются крайне редкими. Для компенсации этих вызовов мы разработали (1) высококачественную систему идентификации языка для более чем 200 языков, которая превосходит общедоступные системы идентификации языка как в наборе данных Flores-200, так и в веб-доменах, (2) детализированную, задокументированную процедуру курирования и очистки одноязычных наборов данных, и (3) методологию обучения мультиязычных кодировщиков предложений на основе учителя-ученика, которая позволяет передачу знаний на языках с крайне ограниченными ресурсами с минимальным количеством контролируемых битекстов. Эти вклады вместе создают более 1.1 миллиарда новых пар предложений для тренировки на 148 языках.

Демократизация создания крупномасштабных наборов данных. Мы предоставляем весь код, который использовался для обучения модели LID и запуска разделения и фильтрации одноязычных предложений. Мы также предоставляем код для майнинга под открытой лицензией, с инструментами для запуска майнинга с использованием наших открытых кодировщиков LASER. Мы создали новую библиотеку для майнинга: stopes24, которая может работать локально или масштабироваться на любом выбранном вами кластере. Эта библиотека поставляется для использования с описанными в этой статье кодировщиками LASER, а также с любыми кодировщиками, доступными в Huggingface sentence transformers.25 Она также позволяет пользователям запускать майнинговый пайплайн end-to-end с помощью простых настраиваемых инструментов. Наша надежда заключается в том, что это позволит большему числу людей создавать высококачественные битекстовые наборы данных на выбранном ими языке и расширять исследования в области языкового майнинга и перевода. Кодировщики можно найти в репозитории LASER, а код учителя-ученика — в ветке nllb репозитория fairseq.



Картинка 14: **Вклады моделирования проекта "No Language Left Behind":** Как показано, мы описываем несколько техник моделирования для обеспечения охвата сотен языков в одной модели. Мы сосредотачиваемся на эффективном масштабировании емкости модели с минимизацией переобучения, а также на улучшении обратного перевода для языков с ограниченными ресурсами и интеграции NLLB-Seed.

6. Моделирование

Существующие исследования в области массового мультиязычного машинного перевода в основном ограничены примерно ста языками, что представляет лишь часть всех письменных языков мира (Arivazhagan et al., 2019; Fan et al., 2020; Zhang et al., 2020). Некоторые работы расширились за эти рамки (Mueller et al., 2020; Siddhant et al., 2022), но критически не имеют надежной оценки производительности. Несмотря на многочисленные исследования, качество перевода для языков с низким объемом данных обычно оставляет желать лучшего. Кроме того, добавление языков с крайне ограниченными ресурсами за пределами ста представляет собой большой вызов из-за часто недостаточного количества качественных битекстов, и даже масштабный майнинг битекстов (Schwenk et al., 2021b) может испытывать трудности в создании достаточно больших наборов данных (см. раздел 5.3). Данные часто значительно варьируются по объему доступной информации на различных языках, что создает дисбаланс между различными языковыми направлениями. Это усложняет обучение массово мультиязычных моделей, поскольку некоторые языковые направления начинают переобучаться, в то время как другие даже еще не сходятся.

В этом разделе мы исследуем, как наиболее эффективно масштабировать мультиязычный машинный перевод до сотен языков. Мы разрабатываем несколько новых техник, которые решают основные проблемы перевода с ограниченными ресурсами, такие как обучение моделей с достаточной емкостью для представления более 200 языков, одновременно адаптируясь к переменной емкости данных для каждой языковой пары. Мы представляем несколько наборов экспериментов, которые помогают нам выявить наиболее производительную архитектуру модели и стратегию обучения, способные масштабироваться до сотен языков (см. раздел 8.1). Эти эксперименты включают (1) модели условных вычислений для минимизации влияния между несвязанными языковыми направлениями, (2) изучение лучших стратегий самообучения на одноязычных данных и (3) аугментацию данных или обратный перевод для направлений с самым низким объемом и низкой точностью.

6.1 Предварительные данные

Сначала мы описываем настройку задачи мультиязычного машинного перевода, включая токенизацию, архитектуру модели и подробности использования набора данных для абляции.

6.1.1 Настройка задачи

Мы моделируем мультиязычный нейронный машинный перевод как задачу последовательность-последовательность, где мы используем кодировщик для условия на входной последовательности на исходном языке и генерируем декодером выходную последовательность в ожидаемом целевом языке (Bahdanau et al., 2015). Мы обучаем модель с максимизацией вероятности перевода на целевом языке T, учитывая предложение на исходном языке S, исходный язык `s и целевой язык `t, то есть P(T|S; `s; `t). В этом разделе мы обсуждаем детали (1) токенизации текстовых последовательностей на исходных и целевых языках, (2) архитектуры модели с особым входом и выходом, специально разработанными для мультиязычного машинного перевода, и (3) набора данных для мультиязычного машинного перевода, который мы используем для экспериментов на меньшем, но репрезентативном масштабе.

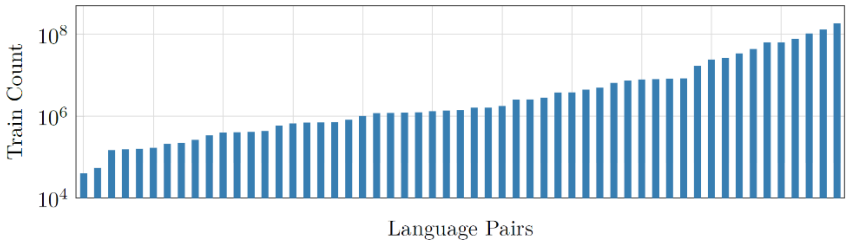
Токенизация с использованием SentencePiece. Для токенизации наших текстовых последовательностей мы обучаем единую модель SentencePiece (SPM) (Kudo и Richardson, 2018) для всех языков. Мы отбираем общее количество 100 миллионов предложений из основных битекстовых данных. Для получения дополнительной информации о основных битекстовых данных см. раздел 8.1.2. Для обеспечения хорошего представительства языков с ограниченными ресурсами в словаре мы уменьшаем выборку высокоресурсных и увеличиваем выборку низкоресурсных языков с температурой выборки 5 (Arivazhagan et al., 2019). Размер словаря является критическим гиперпараметром в моделях мультиязычного перевода с участием языков с ограниченными ресурсами (Oladipo et al., 2022; Rajab, 2022). Размер словаря нашей обученной модели SPM составляет 256,000. Такой большой размер словаря обеспечивает адекватное представление для широкого спектра поддерживаемых языков. Дополнительные подробности см. в разделе 8.1.1.

Архитектура модели. Наша модель мультиязычного машинного перевода последовательность-последовательность основана на архитектуре кодировщик-декодер Transformer (Vaswani et al., 2017). Кодировщик преобразует последовательность токенов исходного языка в последовательность векторных представлений токенов. Декодер использует вывод кодировщика и авторегрессивно генерирует выходную последовательность целевого предложения токен за токеном. Конкретно, кодировщик принимает последовательность токенов W = (w1; : : : ;wS) и исходный язык `s, и генерирует последовательность векторных представлений H = (h1; : : : ; hS), которая затем передается декодеру с целевым языком `t для последовательной генерации целевых токенов V = (v1; : : : ; vT).

$$H = \texttt{encoder}(W,\ \ell_s), \tag{2}$$

$$\forall i \in [1, \ldots, T],\ v_{i+1} = \texttt{decoder}(H,\ \ell_t,\ v_1,\ \ldots,\ v_i). \tag{3}$$

Обратите внимание, что мы добавляем к исходной последовательности исходный язык, в отличие от целевого языка, как это делалось в нескольких работах ранее (Arivazhagan et al., 2019; Johnson et al., 2017). Это в первую очередь потому, что мы приоритизируем оптимизацию нулевой обобщающей способности нашей модели на любую пару из 200 языков за счет небольшого ухудшения надзорной производительности. Эмпирически мы обнаружили, что нулевая обобщающая способность страдает, когда кодировщик настраивается на целевой язык. Когда кодировщик настраивается только на исходный язык, он лучше обобщается на пары исходного и целевого языков, с которыми не сталкивался во время обучения (Fan et al., 2020).



Фигура 15: **Данные абляции**, показывающие количество обучающих данных по всем языковым парам, от 39,992 до 18.7 миллиона предложений.

Как кодировщик, так и декодер представляют собой стеки слоев Transformer. Каждый слой Transformer принимает на вход последовательность вложений и выводит последовательность вложений. В кодировщике слои Transformer состоят из двух подслоев: самовнимания и прямого прохода. Они применяются последовательно и оба предваряются LayerNorm (Ba et al., 2016), за которым следует остаточное соединение (He et al., 2015):

$$Z = X + \texttt{self-attention}(\texttt{norm}(X)), \tag{4}$$

$$Y = Z + \texttt{feed-forward}(\texttt{norm}(Z)). \tag{5}$$

Мы применяем LayerNorm в начале каждого подслоя (Pre-LN), в отличие от его применения после остаточного соединения в конце каждого подслоя (Post-LN). Это делается потому, что в практике Pre-LN оказывается более стабильным по сравнению с Post-LN (Xiong et al., 2020). Слой самовнимания представляет собой слой внимания, который обновляет каждый элемент последовательности, рассматривая другие элементы, в то время как слой прямого прохода (FFN) независимо обрабатывает каждый элемент последовательности через двухслойную MLP. В декодере присутствует дополнительное третье подслое, расположенное между самовниманием и прямым проходом, которое вычисляет внимание к выходу кодировщика. Дополнительные детали можно найти в работе Vaswani et al. (2017).

6.1.2 Датасет абляции

Мы создаем бенчмарк для мультиязычной машинной трансляции, который является представительным для нашего окончательного бенчмарка по более чем 200 языкам. Мы выбираем представительную подвыборку из 53 из 202 языков и общим числом 110 направлений перевода (см. Таблицу 51 в приложении). Среди них 45 направлений относятся к английскому языку (обобщенные как eng_Latn-xx), 45 направлений направлены к английскому языку (обобщенные как xx-eng_Latn) и 20 направлений между не-английскими языками (обобщенные как xx-yy). В зависимости от уровня ресурсов в датасете 40 направлений считаются высокоресурсными, а 70 — низкоресурсными (см. Таблицу 1 и Таблицу 51). Из 70 направлений с низкими ресурсами, 22 являются очень низкоресурсными и содержат менее 100 тыс. обучающих примеров. Датасет состоит из публично доступного битекста для всех 110 языковых направлений (см. Раздел 8.1.2) и масштабированных данных, добытых в большом масштабе (см. Раздел 5.3) для направлений, связанных с английским языком. Всего в этом бенчмарке 864 миллиона примеров. Самая высокоресурсная языковая пара содержит 186 миллионов примеров, а самая низкоресурсная пара — 40 тыс. примеров, что отражает экстремальную несбалансированность в окончательном датасете с 202 языками. На рисунке 15 показано распределение данных по языковым парам, отсортированным по числу примеров на пару. Мы называем этот датасет нашим датасетом абляции и используем его во всех экспериментах данного раздела.

6.2 Условные вычисления для массово-мультиязычного машинного перевода

Модель массово-мультиязычного перевода обучается на нескольких направлениях перевода одновременно, используя общую емкость модели. Это может привести к полезному межъязыковому переносу между связанными языками, но при этом существует риск увеличения взаимного влияния между несвязанными языками (Conneau et al., 2020; Fan et al., 2020). Модели с разреженным использованием смеси экспертов (MoE) являются типом моделей условного вычисления (Almahairi et al., 2016; Bengio et al., 2013), которые активируют подмножество параметров модели на входе, в отличие от плотных моделей, которые активируют все параметры модели на входе. Модели MoE обеспечивают значительную репрезентативную емкость при сохранении той же эффективности вывода и обучения в терминах FLOPs, что и ядреная плотная архитектура. В этом разделе мы изучаем, как можно использовать модели с разреженным использованием смеси экспертов (MoE) (Du et al., 2021; Hwang et al., 2022; Lepikhin et al., 2020; Lewis et al., 2021; Shazeer et al., 2017; Zoph et al., 2022) для достижения более оптимального баланса между межъязыковым переносом и взаимным влиянием, а также для улучшения производительности для языков с низкими ресурсами.

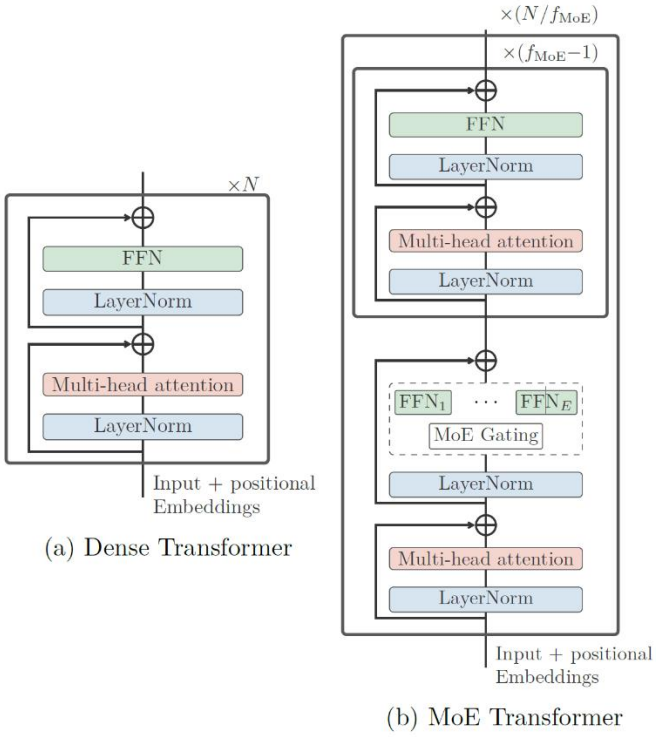
Смесь экспертов с разреженным использованием. Как показано на рисунке 16, мы заменяем подслои прямого прохода (FFN) в плотных моделях подслоем MoE один раз каждые fMoE слоев как в кодировщике, так и в декодере. Подслои MoE состоят из E сетей прямого прохода (FFN), обозначенных как (FFN1; FFN2; ...; FFNE), каждая из которых имеет входные и выходные проекции W(e) i и W(e)o. Каждому подслою MoE прикреплена сеть управления, состоящая из линейного слоя с весами Wg, нормализованного по softmax, чтобы решить, как направлять токены к экспертам. Для данного входного токена xt выход подслоя MoE оценивается как:

$$\mathrm{FFN}_e(x_t) = W_o^{(e)}\,\mathrm{ReLU}(W_i^{(e)} \cdot x_t), \quad (\forall e \in \{1, \ldots, E\}) \tag{6}$$

$$G_t = \mathrm{softmax}(W_g \cdot x_t), \quad \mathcal{G}_t = \text{Top-k-Gating}(G_t), \tag{7}$$

$$\mathrm{MoE}(x_t) = \sum_{e=1}^{E} \mathcal{G}_{te} \cdot \mathrm{FFN}_e(x_t), \tag{8}$$

с вектором маршрутизации Gt 2 RE, вычисляемым сетью управления, то есть для каждого эксперта Gt;e представляет собой вклад eth эксперта (FFNe) в выход MoE. Мы следуем алгоритму Top-k-Gating, предложенному Lepikhin et al. (2020), и направляем каждый токен как минимум к k = 2 экспертам. Мы всегда выбираем двух лучших экспертов для каждого токена, и не добавляем случайности в выбор второго эксперта.



Фигура 16: **Иллюстрация Transformer-кодера с вставленными слоями MoE** с частотой 1:fMoE. Каждый слой MoE содержит E экспертов и сеть управления, ответственную за распределение токенов.

Модель Transformer encoder-decoder, дополненная слоями MoE и соответствующими сетями управления, учится направлять входные токены к двум лучшим экспертам, оптимизируя линейно взвешенную комбинацию сглаженной кросс-энтропии (Szegedy et al., 2015) и вспомогательной потери балансировки нагрузки (Shazeer et al., 2017). Этот дополнительный термин потери (LB) способствует равномерному распределению токенов по экспертам и вычисляется как:

$$LB = E \cdot \sum_{e=1}^{E} f_e p_e, \quad p_e = \frac{1}{T} \cdot \sum_{t=1}^{T} G_{te}, \tag{9}$$

где \(f_e \) - доля токенов, направленных к эксперту \(e \) в качестве первого выбора через Top-k-Gating, а \(p_e \) - средняя вероятность маршрутизации к этому эксперту для \(T \) токенов в мини-пакете. Более подробную информацию по оптимизации моделей MoE можно найти в работе Lepikhin et al. (2020).

В остальной части этого раздела мы сначала подробно описываем, как мы обучаем ванильные модели Sparsely Gated MoE для мультиязычного машинного перевода на нашем бенчмарке и показываем их сравнение с плотными моделями. Затем мы обсуждаем, почему эти ванильные модели Sparsely Gated MoE являются неоптимальными для языков с низким объемом ресурсов (раздел 6.2.1). Мы предлагаем в разделе 6.2.2 ряд архитектурных изменений, которые значительно улучшают производительность на языковых парах с низким объемом ресурсов с использованием моделей MoE. Наконец, мы разрабатываем и изучаем простую, но эффективную стратегию куррикулумного обучения (раздел 6.2.3) как еще один подход для улучшения производительности на языковых парах с низким объемом ресурсов с использованием этих моделей.

	eng_Latn-xx				xx-eng_Latn				xx-yy
	all	high	low	v.low	all	high	low	v.low	all
Dense 615M	41.7	54.2	36.6	30.4	51.1	61.5	46.8	44.0	39.4
MoE-64 (~Dense 615M)	43.0	55.3	37.9	30.3	52.6	63.3	48.3	44.2	39.8
Δ	+1.3	+1.1	+1.3	-0.1	+1.5	+1.8	+1.5	+0.2	+0.4
Dense 615M - p_{drop}=0.1	41.9	54.1	37.0	31.1	51.8	61.8	47.8	45.3	39.6
MoE-64 (~Dense 615M) - p_{drop}=0.1	43.6	55.8	38.7	32.0	53.4	63.6	49.3	45.9	41.1
Δ	+1.7	+1.7	+1.7	+0.9	+1.6	+1.8	+1.5	+0.6	+1.5
Dense 1.3B	43.3	55.4	38.4	31.6	53.5	63.6	49.4	46.5	41.3
MoE-64 (~Dense 1.3B)	43.3	55.9	38.2	29.7	52.9	63.9	48.4	43.7	39.3
Δ	+0.0	+0.5	-0.2	-2.0	-0.6	+0.3	-0.9	-2.8	-2.1
Dense 1.3B - p_{drop}=0.1	43.7	55.4	39.0	33.1	54.4	63.8	50.6	47.9	41.9
MoE-64 (~Dense 1.3B) - p_{drop}=0.3	44.3	56.0	39.5	32.5	54.4	63.9	50.6	47.7	41.9
Δ	+0.6	+0.6	+0.5	-0.6	+0.0	+0.1	+0.0	-0.2	+0.0

Таблица 15: **Ванильные Sparsely Gated MoE с и без общего отсева (набор данных для валидации chrF++).** Мы представляем средние значения для каждого набора направлений: eng_Latn-xx, xx-eng_Latn и xx-yy в целом. Для направлений eng_Latn-xx и xx-eng_Latn мы детализируем пары по уровню ресурсов: высокий ресурс (high), низкий ресурс (low) и очень низкий ресурс (v.low) — Мы видим, что

ванильная модель MoE не превосходит соответствующую плотную модель объемом 1.3B на абляционном бенчмарке. При добавлении общего отсева мы наблюдаем значительное улучшение во всех направлениях на моделях MoE. При меньших вычислительных затратах на обновление (615M), MoE с общим отсевом показывает более значительные приросты.

6.2.1 Ванильные Sparsely Gated MoE и их недостатки для языков с низкими ресурсами

Мотивация использования разреженной активации подсетей экспертов в модели MoE заключается в возможности разделения параметров для моделирования различных аспектов пространства входных данных. Мы предполагаем, что добавленная емкость экспертов должна помочь языковым парам с более высоким уровнем ресурсов, которые в противном случае могли бы ограничиваться использованием одной и той же емкости плотной модели с множеством других языковых пар. Мы также предполагаем, что с огромным количеством направлений перевода добавленная емкость экспертов снизит взаимное влияние, что принесет пользу задачам всех уровней ресурсов. Чтобы проверить этот утверждение и понять пределы ванильных моделей MoE, мы сравниваем в следующем наборе экспериментов производительность моделей MoE с их плотными аналогами на нашем абляционном наборе данных.

Экспериментальная настройка. Мы обучаем базовую плотную модель Transformer encoder-decoder с 1.3 миллиарда параметров: размерность модели 1024, размерность FFN 8192, 16 голов внимания, 24 слоя кодировщика и 24 слоя декодировщика. Затем мы обучаем соответствующую модель Sparsely Gated MoE, заменяя плотный подслои FFN подслоем MoE в каждом втором слое Transformer модели (fMoE=2). Каждый MoE подслои имеет 64 эксперта (близко к числу языков в бенчмарке, то есть 53), и маршрутизирует входные токены к лучшим двум экспертам подслоев FFN в слое MoE, как описано в Lepikhin et al. (2020). Все модели обучаются на 100 тыс. обновлений с эффективным размером пакета 1 миллион токенов на обновление. Для плотных моделей целевая функция - сглаженная кросс-энтропия с параметром сглаживания $\varepsilon = 0.1$ (Szegedy et al., 2015), а для моделей MoE - взвешенная сумма сглаженной кросс-энтропии и потерь балансировки загрузки (Уравнение (9)) с весами 1.0 и 0.01 соответственно. Мы оптимизируем с помощью Adam (Kingma and Ba, 2015) с параметрами $(\beta_1, \beta_2, \varepsilon) = (0.9, 0.98, 10^{-6})$. Мы линейно увеличиваем скорость обучения до 0.004 через 8000 разогревающих обновлений, затем следуем расписанию обратной квадратичной скорости обучения. Для Top-2-Gating мы устанавливаем емкость экспертов равной $2 * T / E$, то есть мы гарантируем, что каждый эксперт обрабатывает максимум $2 * T / E$ токенов, где T - количество токенов в мини-пакете, а E - количество экспертов. Во время генерации мы устанавливаем емкость равной T, чтобы все токены могли быть маршрутизированы к любому выбранному эксперту. Мы используем метрику chrF++ для сравнения производительности моделей (см. раздел 7.1).

Результаты. В таблице 15 мы видим, что модель Sparsely Gated MoE с 64 экспертами (MoE-64), при схожей вычислительной сложности, показывает значительное улучшение по сравнению с плотной моделью на 615M. Мы наблюдаем улучшение оценки chrF++ на 1+ по всем подмножествам, за исключением направлений с очень низким уровнем ресурсов (v.low) и не английских направлений (xx-yy). При увеличении вычислительной сложности на обновление — для модели MoE-64 (сравнимой по вычислительной мощности с плотной моделью на 1.3B), мы наблюдаем нейтральные или худшие результаты по сравнению с плотной моделью на 1.3B. Добавление общего отсева (диапазон параметров pdrop2f0:1; 0:2; 0:3g) значительно улучшает производительность MoE-64 как для варианта с 615M, так и для варианта с 1.3B — для эквивалентного варианта с вычислительной мощностью 615M модель MoE-64 с pdrop=0.1 превосходит плотную модель на 615M с отсевом на +1.5 до +1.7 баллов chrF++ по всем подмножествам направлений. Важно отметить, что при увеличении отсева с 0.0 до 0.1 для MoE-64 мы видим, что относительное снижение -0.1 балла chrF++ переходит в относительное улучшение +0.9 балла chrF++ для направлений с очень низким уровнем ресурсов при переводе из английского. Для эквивалентного варианта с вычислительной мощностью 1.3B мы наблюдаем улучшение производительности на +0.5 до +0.6 баллов chrF++ для направлений с высоким и низким уровнем ресурсов при переводе из английского, но без прироста для направлений в английский, или не английских направлений. Это указывает на то, что при увеличении вычислительной сложности на обновление мы наблюдаем более умеренные улучшения для направлений с высоким уровнем ресурсов, а также для направлений с низким уровнем ресурсов. Мы предполагаем две потенциальные причины этого: (1) мы используем температуру 1.0 для выборки, т.е. мы не увеличиваем выборку данных из направлений с низким уровнем ресурсов. Это сохраняет дисбаланс, который побуждает плотную модель на 1.3B распределять емкость пропорционально уровню ресурсов каждой языковой пары. В результате у языков с высоким уровнем ресурсов вероятно будет достаточно емкости в плотной модели на 1.3B (учитывая размер и характер нашего абляционного набора данных), и они не получат столь значительной выгоды от дополнительной емкости моделей MoE. (2) При увеличении вычислительной сложности на обновление увеличивается склонность к переобучению для направлений с низким или очень низким уровнем ресурсов, что приводит к ухудшению производительности.

Для более глубокого понимания режимов обучения моделей MoE мы анализируем их кривые обучения на рисунке 17. Мы наблюдаем, что в случае направления eng_Latn-kon_Latn, характеризующегося очень низким уровнем ресурсов, модель продолжает сталкиваться с значительным переобучением после обучения на 100 тыс. обновлений. Это неудивительно, так как итерации по небольшому обучающему набору с большой емкостью приводят к переобучению. Важно продолжать обучение для направлений с высоким уровнем ресурсов, но при этом важно избегать негативного влияния на направления с низким уровнем ресурсов. В следующих двух разделах мы обсудим две стратегии, которые решают эту проблему и улучшают производительность моделей на направлениях с низким уровнем ресурсов.

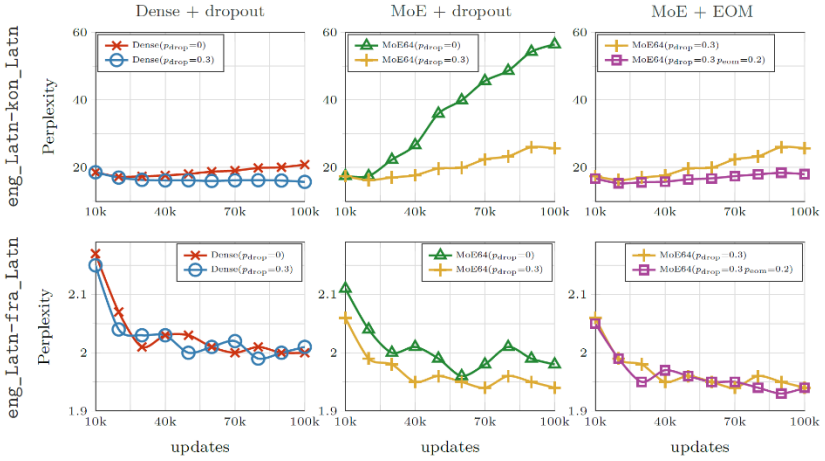


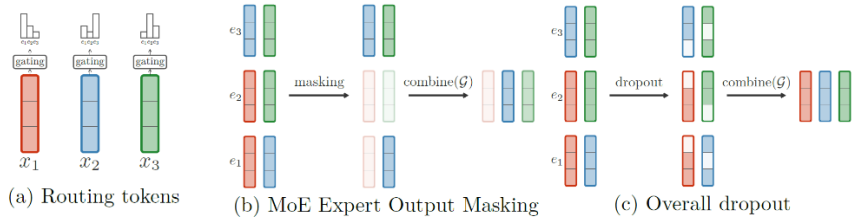
Рисунок 17: **Перплексии на валидации при различных стратегиях отсева** для направления с низким уровнем ресурсов (eng_Latn-kon_Latn в верхнем ряду) и направления с высоким уровнем ресурсов (eng_Latn-fra_Latn в нижнем ряду). Наша предложенная техника маскирования вывода экспертов MoE (EOM) эффективно регуляризирует модель по сравнению с общим отсевом на направлении eng_Latn-kon_Latn.

6.2.2 Регуляризация массово-мультиязычных смесей экспертов

Хотя общий отсев полезен для регуляризации плотных моделей, модели MoE с общим отсевом все еще значительно переобучаются на направлениях с низким уровнем ресурсов, как показано на рисунке 17. Для устранения этого эффекта мы разрабатываем и тестируем различные стратегии регуляризации, специфические для архитектур MoE. Мы описываем каждую из этих стратегий и представляем результаты нашего экспериментального набора данных.

Маскирование вывода экспертов MoE (EOM). Модели MoE позволяют активировать специализированную емкость экспертов на основе входного токена. Однако при увеличении емкости обученное присвоение токенов экспертам может вызывать переобучение моделей, особенно на направлениях перевода с низким уровнем ресурсов. В этой предложенной стратегии регуляризации мы маскируем вывод экспертов для случайной доли (peom) входных токенов. Для входных токенов с отброшенными выводами экспертов первый и/или второй эксперт фактически пропускается. Как показано на рисунке 18b, мы маскируем обоих экспертов для первого токена (x1 красным), не маскируем выводы экспертов для второго токена (x2 синим), и последний сценарий для последнего токена (x3 зеленым), где маскируется только один эксперт. Следует отметить, что хотя это маскирование обнуляет некоторые веса комбинаций Gt;e в уравнении (13), оно не влияет на веса, используемые в потерях балансировки нагрузки в уравнении (9).

Мы сравниваем EOM с Gating Dropout (Liu et al., 2022), стратегией для снижения межмашиноного взаимодействия в слоях MoE, которая также имеет регуляризующий эффект. Gating Dropout пропускает общение All-to-All между GPU с вероятностью pgd, направляя токены вместо этого к локальным экспертам.



Фигура 18: **Иллюстрация маскирования вывода экспертов MoE в отличие от общего отсева для слоев MoE**: каждому цвету соответствует токен, и каждый токен направляется к двум экспертам (Top-k-Gating). Бледные цвета обозначают отброшенные или маскированные выходы.

Final Output Masking (FOM). Проще альтернативой EOM будет маскирование комбинированного вывода экспертов для случайной доли токенов, т.е. после последней стадии на рисунке 18b. Мы обозначаем pfom долю токенов, которые маскируются этим методом регуляризации. Следует отметить, что такой тип маскирования более универсален, так как может применяться и к плотным моделям — его тестирование здесь позволяет подтвердить преимущества использования специфического для MoE маскирования, т.е. Маскирование вывода экспертов MoE.

Conditional MoE Routing (CMR). Вместо случайного отсева части активаций или маскирования вывода экспертов в этом разделе мы рассматриваем возможность позволить модели решать и учить, какие токены нуждаются в дополнительной емкости или специализации слоев MoE, а какие токены лучше направлять в слой общей ограниченной емкости.

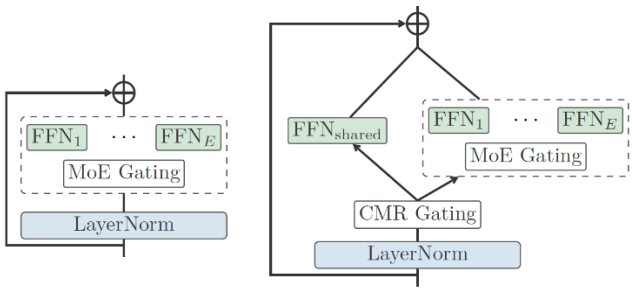
Вдохновленные работой Zhang et al. (2021) по CLSR-Gate, мы проектируем условные слои маршрутизации MoE (CMR). Как показано на рисунке 19, мы расширяем слои MoE двоичным гейтом, который определяет веса, связанные с двумя ветвями вычислительного графа: (1) общий плотный подслой FFN (FFNshared) и (2) слой MoE с собственными подслоями FFN для E экспертов. Для входного токена xt вывод CMR оценивается следующим образом:

$$g(x_t) = \text{sigmoid}(W_{\text{CMR}} \cdot x_t), \tag{10}$$
$$\text{CMR}(x_t) = (1 - g(x_t)) \cdot \text{FFN}_{\text{shared}}(x_t) + g(x_t) \cdot \text{MoE}(x_t), \tag{11}$$

где WCMR — это веса двоичного гейта CMR. В отличие от CLSR-Gate Zhang et al. (2021), наши ветви CMR являются подслоями FFN (плотными или разреженно-групповыми MoE), а не линейными проекциями. Кроме того, наш CMR не имеет языково-специфических параметров, а маршрутизация к экспертам осуществляется с использованием слоя MoE.

Веса гейта CMR обучаются путем оптимизации точности перевода при ограничении бюджета p. Для мини-пакета с T токенами это приводит к добавлению следующего вспомогательного потерянного члена (LCMR) в функцию потерь:

$$L_{\text{CMR}} = \frac{1}{T} \cdot \sum_{t=1}^{T} |g(x_t) - p| \,. \tag{12}$$



Фигура 19: **Условное маршрутизирование MoE (CMR)** - блок с остаточной связью в слое трансформера с обычным MoE (слева) по сравнению с условным маршрутизированием MoE (справа).

Исследование Zhang et al. (2021) показывает, что бюджет p контролирует языко-специфичные возможности — здесь мы используем это ограничение бюджета для управления емкостью, создаваемой слоями MoE; При p=0 модель является плотной, практически направляя все токены через FFNshared, а при p=1 модель свободно маршрутизирует токены через высокоемкостный слой MoE. Это обеспечивает регуляризующий эффект, который может помочь языковым парам с ограниченными ресурсами, склонным к переобучению при использовании больших емкостей слоев MoE. Первоначальные эксперименты на других бенчмарках показали, что значение p=0.8 достигает желаемого баланса между улучшением производительности на языковых парах с ограниченными ресурсами без ущерба для производительности на языковых парах с высокими ресурсами. Важным компонентом, который помогает экспериментальным слоям CMR, является обнуление части ворот g (Уравнение (11)) в мини-пакете; мы обозначаем эту долю как pcmr. Это означает, что мы вынуждаем pcmr% токенов в мини-пакете проходить только через FFNshared. Мы отмечаем, что CMR также связан с работой Rajbhandari et al. (2022), так как оба подхода объединяют выводы плотных и MoE-слоев. Основное отличие заключается в отсутствии вспомогательного бюджетного ограничения и обнулении части ворот g. Из наших первоначальных экспериментов мы находим, что добавление вспомогательного бюджетного ограничения LCMR и pcmr важно для улучшения точности.

Experimental Setup. Наша базовая модель - это модель Sparsely Gated MoE с размерностью модели 1024, размерностью FFN 8192, 16 внимательных головок и 24 слоями. У нас есть 64 эксперта на каждом слое MoE, и мы размещаем слой MoE в каждом втором слое трансформера модели (fMoE=2). Мы сравниваем эту нерегуляризованную базовую модель с вариантом с общей вероятностью отсева 0.3 (pdrop=0.3), лучшей после тщательного анализа pdrop в диапазоне 0.1, 0.2, 0.

- Для EOM мы проходимся по значениям (pdrop, peom) из множества {0.1, 0.2, 0.3} и выбираем лучший из 9 вариантов на основе среднего показателя chrF++ на валидационном наборе.

- Для FOM мы устанавливаем pdrop=0.3 и проходимся по значениям pfom из множества {0.1, 0.2, 0.3}, выбирая лучший вариант на основе среднего показателя chrF++ на валидационном наборе.

- Для CMR мы обучаем базовую модель, дополненную общим FFN-субслоем с теми же размерностями, что и остальная модель. Мы проходимся по значениям (pdrop, pcmr) из множества {0.1, 0.2, 0.3} и выбираем лучший из 9 вариантов на основе среднего показателя chrF++ на валидационном наборе.

- Для Gating Dropout (Liu et al., 2022) мы проходимся по значениям (pdrop, pgd) из множества {0.1, 0.2, 0.3} и выбираем лучший из 9 вариантов на основе среднего показателя chrF++ на валидационном наборе.

	eng_Latn-xx				xx-eng_Latn				xx-yy
	all	high	low	v.low	all	high	low	v.low	all
MoE-64 p_{drop}=0	43.3	55.9	38.2	29.6	52.9	63.9	48.4	43.7	39.3
MoE-64 p_{drop}=0.3 †	44.3	56.0	39.5	32.5	54.4	63.9	50.6	47.7	41.9
MoE-64 FOM (p_{drop}=0.3, p_{fom}=0.1) †	43.8	55.6	38.9	32.5	54.8	64.3	50.9	**48.5**	42.0
MoE-64 EOM (p_{drop}=0.3, p_{eom}=0.1) †	**44.7**	**55.9**	**40.1**	**33.4**	54.8	64.3	51.0	48.3	**42.5**
MoE-64 CMR (p_{drop}=0.2, p_{cmr}=0.2) †	**44.8**	**56.1**	**40.2**	**33.4**	**55.2**	**64.6**	**51.4**	**48.5**	**42.6**
Gating Dropout(p_{drop}=0.3, p_{gd}=0.2) †	44.4	55.7	39.8	33.0	54.8	64.1	51.0	**48.5**	42.3

Таблица 16: **Сравнение различных стратегий регуляризации**. Мы обнаружили, что стратегии EOM и CMR показывают схожие результаты, превосходя базовую модель MoE с общим отсевом. У: лучший результат после анализа.

Для каждой модели мы отчитываемся о средних значениях chrF++ на валидационном наборе данных в трех группах направлений: eng_Latn-xx, xx-eng_Latn и xx-yy, разделенных по уровням ресурсов: высокий, низкий и очень низкий (v.low) для eng_Latn-xx и xx-eng_Latn.

Результаты. С точки зрения снижения проблемы переобучения, последний столбец на рисунке 17 показывает, что EOM обеспечивает лучшую регуляризацию и меньшее переобучение на задачах с низким ресурсом по сравнению с общим отсевом. В отношении качества перевода, как показано в Таблице 16, мы наблюдаем прирост +0.4 chrF++ по всем направлениям в английский и +0.6 chrF++ по направлениям не на английском языке для MoE EOM по сравнению с обычным MoE с общим отсевом. Приросты больше на языках с низким и очень низким ресурсом — для направлений из английского языка есть улучшения +0.6 и 0.9 chrF++ с EOM.

Сравнение между EOM и FOM показывает, что маскирование перед объединением выходов экспертов более выгодно, чем простое маскирование токенов в конечном выводе. Наша гипотеза заключается в том, что этот прирост производительности происходит от укрепления остаточного соединения вокруг слоя MoE и снижения коадаптации между выбранными топ-2 экспертами, а также между экспертами и последующими слоями модели.

Мы обнаружили, что Gating Dropout работает лучше, чем обычный MoE с общим отсевом, но уступает как EOM, так и CMR. Для CMR мы наблюдаем прирост +0.8 chrF++ по всем направлениям в английский и +0.7 chrF++ по направлениям не на английском языке. Аналогичные улучшения наблюдаются для языков с низким и очень низким ресурсом, соответственно +0.7 и +1.0 chrF++.

Эти результаты демонстрируют, что стратегии EOM и CMR помогают улучшить базовый MoE с общим отсевом. CMR требует вычислительных ресурсов больше на 18-25% во время обучения из-за дополнительного общего слоя FFN на уровне каждого слоя MoE в модели. Учитывая дополнительные вычислительные затраты, мы используем более простую стратегию MoE EOM.

6.2.3 Обучение по курсу

Для дальнейшего снижения переобучения на направлениях с низким ресурсом мы исследуем альтернативные методы добавления дополнительной регуляризации. Мы используем простой курс обучения, вводя эти направления с низким ресурсом поэтапно во время обучения модели. Направления, которые эмпирически переобучаются в пределах K обновлений, вводятся за K обновлений до завершения обучения. Это позволяет избежать переобучения направлений, которые склонны к переобучению при слишком большом числе обновлений, и продолжать обучение направлений, которые выигрывают от дополнительных обновлений. Ранее работы исследовали дизайн курса обучения для мультиязычных моделей на основе типов данных (Kuwanto et al., 2021).

Экспериментальная установка. Мы обучаем модели MoE-64 с fMoE=2 в общей сложности T обновлений (см. Раздел 6.2.1 для деталей базовой архитектуры). Для определения фаз курса обучения мы сначала обучаем обычную модель без курса обучения, затем разделяем языковые пары на n корзин fb0; b1; : : : ; bn1g на основе времени начала переобучения. В фазированном обучении по курсу каждую корзину bi вводят после точно T ki обновлений, где ki — это медианный номер обновлений, после которого все направления в корзине bi начинают переобучаться. Исходя из наблюдаемых паттернов переобучения, мы вводим пары во время обучения в n = 3 фазы — мы устанавливаем T = 100k, k0 = 100k, k1 = 40k, k2 = 20k, так что b0 вводится первым, b1 вводится на шаге T 40k, и b2 вводится на шаге T 20k. Мы сравниваем с базовой моделью MoE с общим отсевом 0.3 без обучения по курсу.

Результаты. Как показано в Таблице 17, для обычной MoE при переводе из английского (eng_Latn-xx) наблюдается среднее улучшение +0.6 chrF++ на направлениях с низким ресурсом (low) и улучшение +0.8 chrF++ на направлениях с очень низким ресурсом (v.low). Однако значительного улучшения нет на направлениях с высоким ресурсом (high) или при переводе на английский язык (xx-eng_Latn), скорее всего, потому что в базовой модели на этих направлениях нет переобучения. Для MoE EOM обучение по курсу фактически вредит производительности на направлениях eng_Latn-xx, xx-eng_Latn и xx-yy. Наш анализ указывает на то, что переобучение на абляционном наборе данных уже снижено с помощью EOM, который имеет больший процент языковых пар с лучшим контрольным показателем по perplexity на шаге 70000. Мы предполагаем, что добавление курса обучения поверх EOM не нужно для абляционного набора данных. Однако результаты в Таблице 29 показывают, что с полным набором данных и более крупной моделью комбинация обучения по курсу и EOM улучшает производительность, особенно на направлениях с низким и очень низким ресурсом.

6.2.4 Анализ мультиязычных моделей с разреженными MoE

Теоретически MoE позволяет моделям специализировать экспертные возможности для различных задач, но что на самом деле учат эти модели? Мы теперь более внимательно рассмотрим маршрутизацию токенов к экспертам в слоях MoE в различных точках архитектуры энкодера-декодера. Мы берем модель MoE (E=64, pdrop=0:3, peom=0:2), обученную на нашем абляционном наборе

данных, и выполняем проход вперед на наборе данных dev Flores-200 в режиме преподавания учителя, т.е. мы подаем истинный префикс целевого токена для предсказания следующего целевого токена. Для каждой задачи (языковая пара) мы регистрируем решения о маршрутизации до Top-k-Gating, и в зависимости от того, является ли это энкодерным слоем или декодерным слоем, мы усредняем векторы маршрутизации по нескольким языковым парам для оценки языковых векторов маршрутизации.

	eng_Latn-xx				xx-eng_Latn				xx-yy
	all	high	low	v.low	all	high	low	v.low	all
MoE-64 (p_{drop}=0.3)	44.3	56.0	39.5	32.5	54.4	63.9	50.6	47.7	41.9
+ Curriculum Learning (3 Phases)	44.7	56.0	40.1	33.3	54.6	64.2	50.8	47.9	42.2
Δ	+0.4	+0	+0.6	+0.8	+0.2	+0.3	+0.2	+0.2	+0.3
MoE-64 EOM (p_{drop}=0.3, p_{eom}=0.1)	44.7	55.9	40.1	33.4	54.8	64.3	51.0	48.3	42.5
+Curriculum Learning (3 Phases)	44.26	55.68	39.62	33.07	54.67	63.87	50.93	48.44	42.15
Δ	-0.44	-0.22	-0.48	-0.33	-0.13	-0.43	-0.07	+0.14	-0.35

Таблица 17: **Результаты обучения по курсу** показывают, что для обычной MoE обучение по курсу снижает переобучение, особенно для направлений eng_Latn-xx с низким и очень низким ресурсом. Для MoE EOM обучение по курсу не помогает.

$$G_{<\text{lang}>} = \frac{1}{|\mathcal{T}_{<\text{lang}>}|} \sum_{x_t \in \mathcal{T}_{<\text{lang}>}} G_t, \qquad (13)$$

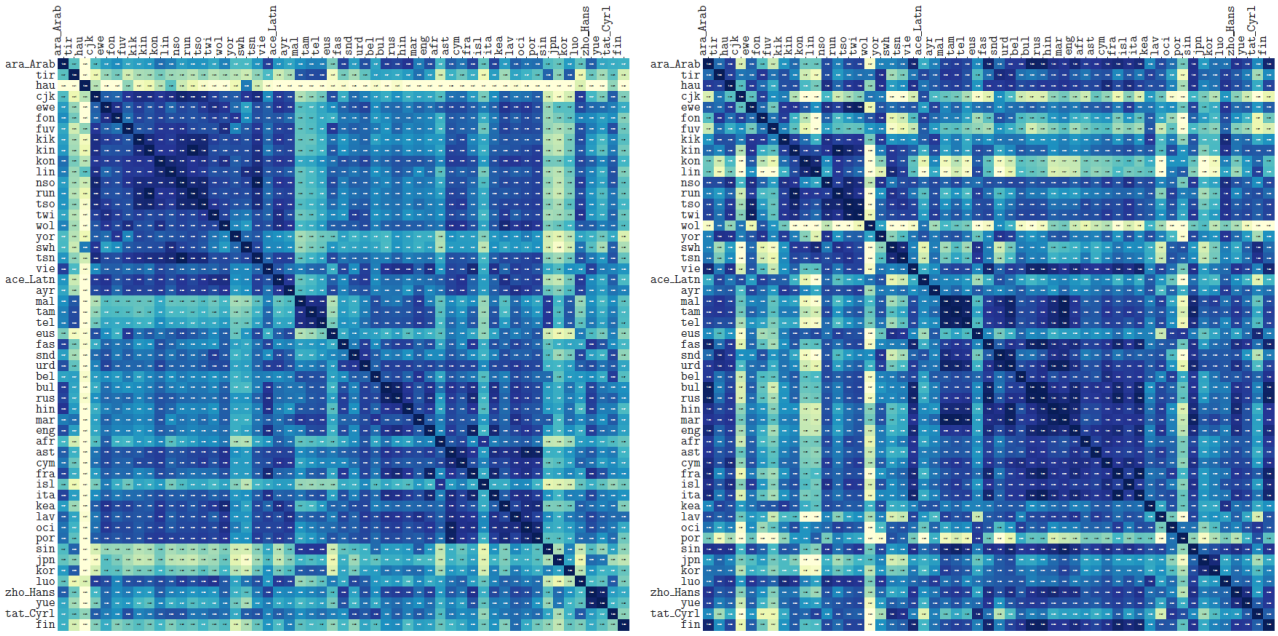
где T<lang> — это набор всех токенов в <lang>, на стороне источника для слоев энкодера и на стороне цели для слоев декодера. Мы строим на Рисунке 20 график косинусных коэффициентов сходства между всеми 53 языками набора данных абляции в первом и последнем слое MoE энкодера, а также в первом и последнем слое MoE декодера.

Тепловые карты сходства демонстрируют, что в поздних слоях декодера (см. Рисунок 20d) языки разделяются, т.е. направляются к разным наборам экспертов. Языки в пределах одной семьи имеют высокую степень сходства в выборе экспертов, т.е. поздние слои MoE декодера являются специфичными для языка. Это особенно верно для языков Атлантико-Конголезской семьи (строки/столбцы от cjk до yor) и некоторых пар, таких как fsnd_Arab; urd_Arabg в Индо-Европейской семье или fyue_Hant; zho_Hansg в Сино-Тибетской семье. В меньшей степени ранние слои MoE энкодера (см. Рисунок 20a) также показывают некоторую специализацию по языкам и экспертам. Поздние слои MoE энкодера и ранние слои MoE декодера (см. Рисунок 20b и Рисунок 20c) кажутся нейтральными к языкам.

На Рисунке 21 мы визуализируем векторы распределения экспертов по языкам G<lang> (Уравнение (13)) с использованием UMAP (McInnes et al., 2018). В первой строке отображены языки по языковой семье, а во второй строке — языки по алфавиту. Разделение языков более заметно в последнем слое декодера (последний столбец Рисунка 21), особенно по линиям языковых семей, например, Атлантико-Конго зелёным цветом и Дравидийская семья розовым цветом.

6.3 Стратегии самообучения на крупных монолингвальных корпусах

Для языков с низкими ресурсами обычно доступно ограниченное или вовсе недоступное количество двуязычных данных. В случаях, когда двуязычные данные общедоступны, их область применения может быть узкой (например, религиозные тексты) или данные могут быть шумными. По сравнению с этим, монолингвальные данные в низкоресурсных языках доступны в более значительных объемах. В дополнение к майнингу двуязычных текстов, описанному в Разделе 5.3, ещё одним способом использования этих монолингвальных данных является включение задачи самообучения в процесс обучения многозадачной машинной модели перевода (Bapna et al., 2022; Chi et al., 2021; Ma et al., 2021). Самообучение может изучать модели и структуры языка из монолингвального текста. Следовательно, обучение с самообучающимися целями на этих дополнительных данных может помочь улучшить производительность моделей многоязычного машинного перевода.



(a) First encoder MoE layer (b) Last encoder MoE layer

(c) First decoder MoE layer (d) Last decoder MoE layer

Рисунок 20: **Косинусное сходство** между языками из абляционного набора данных на разных слоях архитектуры кодировщика-декодера. Сходство измеряется относительно решений о выборе экспертов для каждого языка (на стороне источника в кодировщике и на стороне цели в декодере).

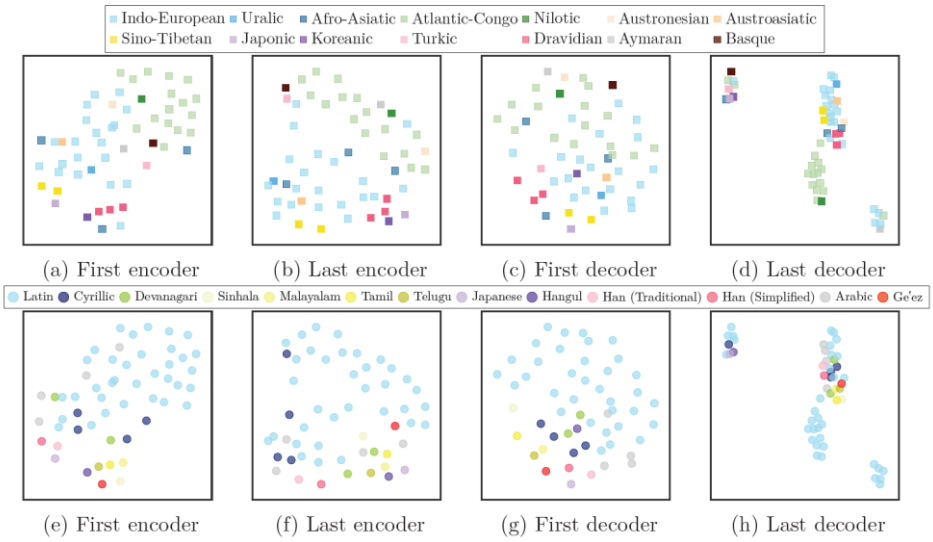


Рисунок 21: **Встраивание UMAP языков в абляционном наборе данных**. В первой строке используется цветовое обозначение по языковой семье, во второй строке - по письменности.

Сначала мы изучаем эффект выбора задачи самообучения: (1) задача языкового моделирования, (2) задача денойзинг автокодировщика (Liu et al., 2020) или (3) их комбинация. Затем мы изучаем оптимальную стратегию комбинирования задачи самообучения (SSL) на монолингвальных данных с задачей обучения на многоязычном машинном переводе (MMT) на парных текстах. Мы используем рекомендации этого исследования для выбора стратегии самообучения для нашей финальной модели на 200 языках.

6.3.1 Включение целей самообучения в многоязычный машинный перевод

Существуют различные способы включения самообучения на монолингвальных данных в процесс обучения моделей многоязычного машинного перевода. Традиционно, самообучение в NLP принимает форму предобучения с целью самообучения на монолингвальных данных, а затем последующего тонкого настройки на задачно-специфических данных с учителем. Другая стратегия заключается в том, чтобы рассматривать SSL на монолингвальных данных и MMT на парных текстах как отдельные задачи в установке многозадачного обучения, где примеры из обеих задач присутствуют в каждом батче во время обучения. Наша третья стратегия - это комбинация первых двух стратегий. Сначала мы можем предобучить модель в многозадачной установке с самообучением на монолингвальных данных и многоязычным машинным переводом на парных текстах, а затем выполнять тонкую настройку только на многоязычном машинном переводе.

6.3.2 Цели самообучения

Существуют различные цели самообучения, которые мы можем использовать на монолингвальных данных.

Денойзинг автокодировщик (DAE).** Мы следуем Liu et al. (2020) и используем архитектуру Transformer encoder-decoder, которая также используется для многоязычного машинного перевода. Мы задаем цель как предложение из монолингвального корпуса. Источником является версия целевого монолингвального предложения с шумом. Мы маскируем случайные фрагменты текста из целевого предложения. Вместо того, чтобы всегда заменять замаскированный токен на токен <mask>, с определенной вероятностью мы заменяем замаскированный токен на случайный токен из словаря. Цель SSL здесь - максимизировать вероятность предсказания цели, учитывая источник, который является зашумленной версией цели.

Казуальное языковое моделирование (LM). Прошлые исследования показали некоторый успех в инициализации компонентов модели машинного перевода с предварительно обученной языковой моделью Transformer декодера. Поэтому мы устанавливаем задачу языкового моделирования в seq2seq установке, где источник пуст, а целью является предложение из монолингвального корпуса. Цель SSL здесь - максимизировать вероятность предсказания цели, т.е. цель казуального языкового моделирования.

6.3.3 Влияние учебного плана самообучения в сочетании с многоязычным машинным переводом

В этом разделе мы изучаем влияние нашего выбора учебного плана при комбинировании цели денойзинг автокодировщика (DAE) с обычной задачей MMT.

Экспериментальная установка. В нашей установке мы используем цель многоязычного денойзинг автокодировщика для самообучения на многоязычном текстовом корпусе для языков в нашем контрольном наборе. Нашей базовой моделью является плотная модель на 1.3B, обученная на задаче MMT в течение 100K обновлений с температурой выборки 1.0. Мы обучаем варианты задачи SSL с той же архитектурой и температурой выборки в течение 200K обновлений, чтобы гарантировать, что варианты задачи SSL будут подвергаться воздействию такого же количества многоязычных параллельных данных, как и базовая модель. Мы надеемся найти лучший вариант, который улучшит производительность на парах с низкими ресурсами, у которых нет достаточного количества параллельных битекстов, но относительно много монолингвальных данных. Мы рассматриваем три варианта:

- **Предобучение на SSL, за которым следует тонкая настройка на MMT (DAE)MMT).** Мы обучаем на SSL первые 100K обновлений, после чего выполняем тонкую настройку на задаче MMT следующие 100K обновлений. Мы используем температуру выборки 1.25 для предобучения SSL.

- **Многозадачное обучение на SSL и MMT (DAE+MMT).** Мы объединяем примеры обучения для SSL и MMT и обучаем модель в течение 200K обновлений на комбинации обеих задач.

- **Многозадачное обучение на SSL и MMT, за которым следует тонкая настройка на MMT (DAE+MMT) MMT)**. Сначала мы выполняем многозадачное обучение на SSL и MMT в течение 150K обновлений, а затем выполняем тонкую настройку только на MMT в течение 50K обновлений.

Результаты. Во-первых, мы видим, что предобучение на SSL, за которым следует тонкая настройка на MMT (DAE → MMT), ухудшает производительность (-0.7 chrF++ на eng_Latn-xx, -1.2 chrF++ на xx-eng_Latn) во всех подмножествах пар по сравнению с базовой моделью, обученной только на MMT.

	eng_Latn-xx				xx-eng_Latn				xx-yy
	all	high	low	v.low	all	high	low	v.low	all
MMT	43.3	**55.4**	38.4	31.6	53.5	**63.6**	49.4	46.5	41.3
DAE⇒MMT	42.6	55.0	37.6	30.8	52.3	62.2	48.3	45.4	40.4
DAE+MMT	**43.5**	55.2	**38.8**	**32.7**	**54.4**	**63.6**	**50.7**	**48.4**	**42.4**
DAE+MMT⇒MMT	43.4	**55.4**	38.5	32.2	54.3	**63.6**	50.5	48.0	42.2

Таблица 18: **Эффект использования плана обучения SSL**. Мы обнаружили, что совместное обучение DAE+MMT является самой эффективной стратегией.

Это указывает на фундаментальное несоответствие между двумя задачами обучения, и сложно восстановиться в процессе тонкой настройки после начального состояния в конце предварительного обучения. Многозадачное обучение на SSL и MMT (DAE+MMT) показывает +0.4 chrF++ на парах с низкими ресурсами в направлении eng_Latn-xx и +1.1 chrF++ на парах с очень низкими ресурсами в том же направлении, что подтверждает полезность дополнительного монолингвального материала на языках с низкими ресурсами при его представлении в многозадачной структуре. Аналогично, мы видим улучшения на +1.3 и +1.9 chrF++ на парах с низкими и очень низкими ресурсами в направлении xx-eng_Latn, а также +1.1 chrF++ на парах xx-yy. Мы предполагаем, что наблюдаем более сильную производительность при переводе на английский язык, так как имеется огромное количество очень качественного монолингвального материала на английском языке, что критично для SSL. Наконец, мы хотим подтвердить, есть ли какие-либо преимущества в приостановке многозадачного обучения после первых K обновлений и последующей тонкой настройке только на MMT (DAE+MMT)MMT. Мы видим, что это выполняется на уровне с многозадачным обучением на SSL и MMT (DAE+MMT) без тонкой настройки только на MMT. Это указывает на то, что когда SSL и MMT совместно обучаются в многозадачном режиме, они не страдают от интерференции, которая могла бы быть устранена на последнем этапе тонкой настройки исключительно на окончательной задаче (MMT).

6.3.4 Влияние целей самоконтроля

В этом разделе мы сравниваем влияние выбора различных целей самоконтроля.

Экспериментальная настройка. Мы обучаем все варианты задач SSL в этом разделе в многозадачном режиме обучения задачам SSL и MMT вместе, потому что это наилучший план, как показано в разделе 6.3.3. Наша базовая модель - плотная модель на 1.3B, обученная на 100K обновлений, а модели SSL - на 200K обновлений. Мы обучаем варианты SSL с той же архитектурой и температурой выборки в течение 200K обновлений каждая, чтобы убедиться, что варианты SSL подвергаются воздействию такого же количества обучающих данных, как и базовая модель. Мы надеемся найти наилучшую цель самоконтроля, которая улучшит производительность на парах с низкими ресурсами, у которых недостаточно данных bitext, и относительно обильные монолингвальные данные. Мы рассматриваем три варианта: (1) Декодер автоэнкодера (DAE), (2) Языковое моделирование (LM) и (3) Комбинацию обоих (DAE+LM).

Результаты. Во-первых, мы наблюдаем, что задача LM в качестве цели самоконтроля приводит к снижению производительности по сравнению с базовой моделью, обученной только на MMT. Одно из предположений состоит в том, что самоконтроль на внимании между энкодером и декодером играет важную роль, и это не присутствует в цели LM, которая сосредоточена на обучении декодера для генерации текста в монолингвальном корпусе.

	eng_Latn-xx				xx-eng_Latn				xx-yy
	all	high	low	v.low	all	high	low	v.low	all
MMT	43.3	**55.4**	38.4	31.6	53.5	**63.6**	49.4	46.5	41.3
MMT+LM	42.6	54.9	37.5	30.8	53.5	**63.6**	49.4	46.7	41.5
MMT+DAE	**43.5**	55.2	**38.8**	**32.7**	**54.4**	**63.6**	**50.7**	**48.4**	**42.4**
MMT+DAE+LM	42.6	55.0	37.6	31.4	53.4	62.7	49.6	47.0	40.8

Таблица 19: Влияние различных целей самоконтроля. Мы считаем, что обучение MMT+DAE является наиболее эффективным по сравнению с добавлением задачи LM.

Далее, как также показано в разделе 6.3.3, мы видим значительные улучшения при использовании цели DAE. Наконец, мы исследуем, есть ли дополнительные преимущества от комбинирования DAE с задачей LM. Результаты показывают снижение производительности по сравнению с использованием только задачи DAE. Это предполагает, что может существовать некоторая интерференция между различными задачами, что снижает общую производительность при их совмещении.

6.3.5 Обсуждение

Недавние работы (Bapna et al., 2022; Chi et al., 2021; Ma et al., 2021) показали, что денойзинг и аналогичные самоконтролируемые цели очень полезны для улучшения производительности модели, когда они обучаются вместе с задачей машинного перевода в мультимасштабной настройке. В нашей работе мы пробуем две цели SSL: DAE и LM, и экспериментируем с различными комбинациями обеих вместе с задачей MMT. Мы обнаружили, что только DAE показывает наилучшие результаты при обучении с MMT. Преимущества задачи LM в мультимасштабной настройке с MMT пока не очевидны, и будущие исследования могут дать более глубокое понимание этого вывода. Мы также изучаем различные стратегии обучения по учебному плану с задачами SSL и обнаруживаем, что мультимасштабное обучение DAE и MMT обычно является наилучшей настройкой, что аналогично выводам в (Chi et al., 2021; Ma et al., 2021). В разделе 8.1.2 далее обсуждается, как мультимасштабное обучение SSL+MMT улучшает производительность модели для создания более качественных данных с обратным переводом. Самоконтролируемое обучение — это мощная техника для оптимального использования монолингвальных данных, и есть возможность изучать и разрабатывать лучшие цели SSL для этого. Кроме того, разработка учебных планов для комбинирования этих целей в мультимасштабной структуре также представляет интересное направление для будущих исследований.

6.4 Увеличение данных

Еще один способ использования монолингвальных данных — это обратный перевод (Edunov et al., 2018; Sennrich et al., 2016a), техника, которая включает создание параллельных корпусов, содержащих шумные данные на исходной стороне с помощью машинного перевода. Однако, когда речь идет о языках с ограниченными ресурсами, модели машинного перевода, используемые для создания данных обратного перевода, часто недостаточно хороши, и поэтому сгенерированные данные часто оказываются шумными и вырожденными. Hoang et al. (2018) предложили итеративный обратный перевод для компенсации этого, так как для каждой итерации используются более лучшие модели для генерации обратных переводов. Однако это дорогостоящее занятие. Поскольку цель состоит в том, чтобы использовать наилучшие возможные модели для создания данных обратного перевода, это означает использование масштабных многоязычных моделей, которые являются вычислительно интенсивными как на этапе обучения, так и на этапе вывода. В этом разделе мы обсуждаем, как мы можем эффективно генерировать высококачественные данные обратного перевода для языков с ограниченными ресурсами, используя одну итерацию для повышения вычислительной эффективности.

6.4.1 Разные источники данных

Обратный перевод является источником синтетически дополненных данных для моделей перевода. Мы сравниваем его с двумя другими источниками данных: (1) основными битекстовыми данными, то есть высококачественными параллельными корпусами, переведенными человеком, и (2) добытыми битекстовыми данными, то есть параллельными корпусами, полученными путем майнинга (см. раздел 5.3). Обратный перевод зависит от доступности начальной системы перевода, учителя, для создания шумного параллельного корпуса из монолингвального корпуса. Естественный выбор для системы учителя — использование модели на основе трансформеров, аналогичной нашим предыдущим подходам. Однако эти нейронные системы часто считаются неэффективными с точки зрения данных по сравнению с традиционными фразовыми или основанными на правилах статистическими моделями машинного перевода (Koehn и Knowles, 2017). В свете этого мы также экспериментируем с использованием традиционной фразовой статистической модели перевода в качестве учителя для обратного перевода (Schwenk, 2008).

Таким образом, в следующих экспериментах мы будем различать четыре различных источника данных: один основной и три дополненных источника: данные, переведенные человеком; добытые данные; данные обратного перевода через многоязычную нейронную модель машинного перевода; и данные обратного перевода через статистическую модель машинного перевода. Для краткости и дальнейшего использования мы показываем характеристики этих источников данных в таблице 20 и используем следующие сокращенные термины для источников данных:

- **NLLB-Seed**: наши профессионально переведенные начальные наборы данных, описанные в разделе 4.2.

- **PublicBitext**: общедоступные параллельные корпуса. Эти наборы данных могут быть переведены человеком, но часто автоматически выравнены.

- **Primary**: комбинация двух вышеуказанных источников.

- Mined: добытые данные, как описано в разделе 5.3.

- **MmtBT**: данные обратного перевода, полученные с помощью плотной многоязычной нейронной модели с 1,3 миллиардами параметров.

- **SmtBT**: данные обратного перевода, полученные с помощью серии двуязычных моделей MOSES (Koehn et al., 2007), обученных на Primary и Mined данных. Оптимальные гиперпараметры модели были выбраны с помощью валидационных данных Flores-200.

Экспериментальная установка.** Сначала мы изучаем эффект использования MmtBT в дополнение к Primary и Mined данным. Мы используем плотную модель с 1,3 миллиардами параметров и обучаем ее на двух наборах данных: Primary и Primary+Mined. Затем мы используем модель, обученную на наборе данных Primary+Mined, и генерируем данные обратного перевода для всех пар, ориентированных на английский язык, в наборе данных. Эти данные обратного перевода становятся нашим MmtBT. В следующем

эксперименте мы используем точно такую же модель и обучаем ее на наборе данных, включающем Primary+Mined+MmtBT. Наша цель здесь — наблюдать преимущества обратного перевода по сравнению с Primary+Mined. Мы обучаем все модели на протяжении 200 000 обновлений и сравниваем лучшие контрольные точки на наборе данных для разработки Flores-200, используя метрику chrF++.

Source	Human Aligned?	Noisy?	Limited Size?	Model-Dependent?	Models Used
NLLB-Seed	✓	✗	✓	✗	—
PublicBitext	✗	✓	✓	✗	—
Mined	✗	✓	✗	✓	Sentence Encoders
MmtBT	✗	✓	✗	✓	Multilingual
SmtBT	✗	✓	✗	✓	Bilingual MOSES
Ideal Data	✓	✗	✗	✗	—

Таблица 20: **Характеристики наборов** данных источников, с которыми мы сравниваемся в этом разделе. Из этих наборов данных NLLB-Seed является самым маленьким. Для языков с ограниченными ресурсами PublicBitext часто является крайне ограниченным. Mined, MmtBT и SmtBT ограничены только объемом доступных одноязычных данных и качеством моделей, используемых для их создания.

	eng_Latn-xx				xx-eng_Latn				xx-yy
	all	high	low	v.low	all	high	low	v.low	all
Primary	41.0	52.8	36.3	28.1	47.4	60.5	42.1	36.7	39.2
+Mined	43.8	55.2	39.2	**34.0**	53.9	64.4	49.6	46.1	40.9
+MmtBT	44.0	55.1	39.5	**34.0**	55.7	64.8	52.0	50.8	40.6
+SmtBT	**44.2**	**55.5**	**39.6**	**34.0**	**55.9**	**64.9**	**52.2**	**50.9**	**41.1**

Таблица 21: **Сравнение общей производительности модели, обученной на различных комбинациях данных**, оцененная на наборе данных Flores-200 для направлений абляционного набора. Мы наблюдаем, что добавление данных SmtBT улучшает результаты по сравнению с Mined+MmtBT и в целом обеспечивает лучшую производительность по всем направлениям языков и типам уровня ресурсов.

Во втором наборе экспериментов мы пытаемся понять выгоды добавления дополнительного источника обратно переведённых данных, SmtBT. Для этого мы обучаем двуязычные статистические модели машинного перевода (SMT) на первичных и добытых битекстах. Мы сравниваем производительность этих моделей с моделью мультиязычного машинного перевода (MMT), обученной на том же наборе данных, и выбираем направления, где модели SMT либо лучше, либо сопоставимы с моделями MMT. Для выбранных направлений мы генерируем обратно переведённые данные, используя модели-учителя SMT. Это даёт нам набор данных SmtBT. Для сравнения дополнительных выгод SmtBT мы объединяем все источники данных: Primary+Mined+MmtBT+SmtBT, обучаем аналогичную модель с плотным слоем 1.3B и сравниваем её производительность.

Результаты. В Таблице 21 мы отчитываемся о производительности базовой модели, использующей только данные Primary, а затем о трёх других моделях, обученных путём постепенного дополнения тренировочных данных наборами Mined, MmtBT и SmtBT. Мы видим, что наивысшая производительность достигается при использовании всех источников данных. Несмотря на недавние достижения, которые подвергли сомнению предполагаемую неэффективность данных в нейронном машинном переводе (Sennrich and Zhang, 2019), мы видим, что использование SMT в качестве источника обратного перевода всё ещё приводит к улучшениям для направлений с очень низкими ресурсами. Мы можем дальше исследовать качество обратного перевода, рассматривая производительность наших двух моделей-учителей: модели MMT и набора двуязычных моделей SMT, обученных на том же наборе данных Primary+Mined.

Method	avg chrF++	#best
SMT	23.1	8
MMT	**43.7**	**30**

(a) eng_Latn-xx

Method	avg chrF++	#best
SMT	26.2	0
MMT	**54.8**	**38**

(b) xx-eng_Latn

В таблице 22 модель MMT, как ожидалось, превосходит традиционные SMT почти во всех направлениях. Несмотря на то что средняя производительность SMT может быть низкой, мы предполагаем, что добавление различных, взаимодополняющих источников шума является причиной, почему его дополнение всё ещё полезно для общей производительности модели. Различия между поколениями можно визуализировать по частоте токенов, сгенерированных моделями MMT и SMT при переводе одного и того же корпуса. Для примера гистограмм см. на рисунке 42 в Приложении D.

6.4.2 Маркировка данных

Маркированный обратный перевод (Caswell et al., 2019) — это техника, которая помогает модели различать различные источники данных, с которыми она сталкивается во время обучения. Это достигается путём добавления специальных токенов к обратно переведённым обучающим примерам и показано, что это повышает производительность, помогая модели различать шумные данные и избегать их переобучения (Marie et al., 2020). В экспериментах предыдущего раздела мы использовали расширенную схему маркировки, где использовались специальные токены для каждого из трёх источников данных: <MINED_DATA> для Mined, <MMT_BT_DATA> для MmtBT и <SMT_BT_DATA> для SmtBT. Мы изучаем эффекты отсечения этой схемы маркировки, экспериментируя с использованием одной метки для всех вторичных данных (<SECONDARY_DATA>), а также без использования меток вообще.

Экспериментальная установка

Мы обучаем модели MMT на полном наборе данных, состоящем из Primary, Mined, MmtBT и SmtBT, но убираем различные схемы маркировки. В настройке без меток не используются метки вообще, в настройке с одной меткой используется одна и та же метка для Mined, MmtBT и SmtBT. Наконец, в настройке с тонкозернистыми метками используются отдельные метки для Mined, MmtBT и SmtBT.

Результаты

Результаты в таблице 23 демонстрируют преимущества использования тонкозернистых меток. Это дополнительно подтверждает гипотезу Caswell et al. (2019) о том, что маркировка полезна для того, чтобы помочь модели различать синтетические и естественные данные. Это также подтверждает, что указание на специфический характер синтетических данных может дополнительно повысить производительность.

6.5 Начальное формирование моделей с использованием NLLB-Seed

Для значительного числа низкоресурсных языков, рассмотренных в данной работе, параллельные корпуса, доступные для исследований, часто состоят всего из нескольких тысяч предложений. Они часто поступают из источников с очень узкой спецификой, таких как религиозные тексты, и уровень контроля качества часто неясен. В то время как перевод миллионов предложений профессиональными переводчиками невозможен, перевод нескольких тысяч предложений вполне выполним.

	eng_Latn-xx				xx-eng_Latn				xx-yy
	all	high	low	v.low	all	high	low	v.low	all
No Tags	42.8	54.5	38.0	31.9	54.8	64.2	50.9	48.4	40.8
Single Tag	44.0	55.2	39.4	**34.2**	55.5	64.6	51.8	50.5	40.7
Finegrained Tags	**44.2**	**55.5**	**39.6**	34.0	**55.9**	**64.9**	**52.2**	**50.9**	**41.1**

Таблица 23: **Сравнение различных схем маркировки** на наборе данных Flores-200 для проверочного набора. Мы сравниваем модели, обученные на наборе данных Flores-200, используя отсутствие маркеров, один маркер и мелкозернистые маркеры. Мы сообщаем оценки chrF++, агрегированные по направлению языка и уровню ресурсов. Мы замечаем, что мелкозернистая маркировка дает лучшую производительность.

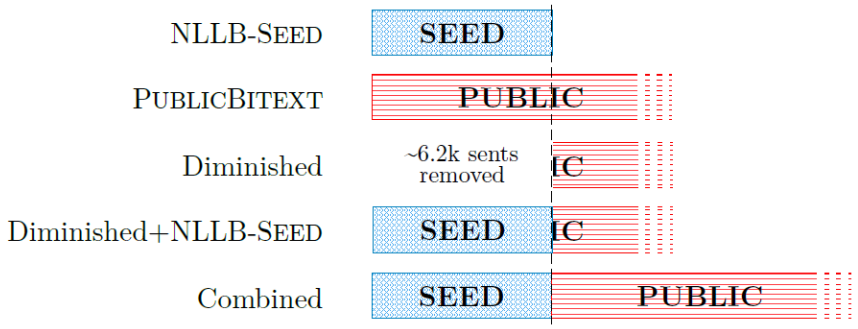


Рисунок 22: **Конфигурации набора данных** для абляционных экспериментов раздела 6.5. Мы измеряем влияние использования NLLB-Seed в различных экспериментальных условиях.

Важно понять, есть ли ценность в таких маленьких, но высококачественных человекоаннотированных исходных наборах данных для языков с ограниченными ресурсами. Достаточно ли такого маленького набора данных для запуска системы машинного перевода для нового языка с ограниченными ресурсами или для настройки существующей системы машинного перевода на новый домен? В этом разделе мы исследуем эти вопросы и стремимся измерить эффект обучения моделей перевода на таком небольшом человекоаннотированном исходном наборе данных, как NLLB-Seed (см. Раздел 4.2). Нас интересует количественная оценка важности набора данных, который был профессионально переведен, охватывает более широкий домен и может быть уверенно отнесен к указанному языку.

6.5.1 Полезность NLLB-Seed

Во-первых, мы хотим измерить производительность двуязычных моделей, обученных на NLLB-Seed, по сравнению с теми, которые были обучены на общедоступных данных (PublicBitext), а также на их комбинации (NLLB-Seed+PublicBitext). Во-вторых, имея определенное количество общедоступных параллельных данных, мы изучаем инкрементный эффект добавления NLLB-Seed по сравнению с простым добавлением большего количества данных из первоначального домена общедоступного битекста. Чтобы ответить на это, мы создаем для каждого языка новый набор данных, который мы называем Diminished+NLLB-Seed. Этот набор данных получен путем вычитания из PublicBitext случайной выборки предложений того же размера (≈6,2 тыс.), а затем замены их на NLLB-Seed. Результатом является набор данных того же размера, что и PublicBitext, но который также содержит NLLB-Seed. Составы наборов данных, которые мы изучаем здесь, изображены на рисунке 22.

Direction	PublicBitext #train	PublicBitext chrF++	NLLB-Seed #train	NLLB-Seed chrF++	Combined chrF++
Average		14.5		17.4	20.5
ban_Latn-eng_Latn	10.2 k	13.1	6.2 k	20.8	**22.2**
eng_Latn-ban_Latn	10.2 k	15.9	6.2 k	20.6	**21.9**
dik_Latn-eng_Latn	16.9 k	12.9	6.2 k	16.1	**17.0**
eng_Latn-dik_Latn	16.9 k	9.0	6.2 k	**13.7**	13.1
fuv_Latn-eng_Latn	12.1 k	15.6	6.2 k	16.3	**18.1**
eng_Latn-fuv_Latn	12.1 k	9.2	6.2 k	9.8	**13.5**
mri_Latn-eng_Latn	31.3 k	16.7	6.2 k	17.4	**26.8**
eng_Latn-mri_Latn	31.3 k	23.2	6.2 k	24.3	**31.5**

Таблица 24: **Размер обучающего набора данных и производительность валидации FLORES-200 для двуязычных моделей с ограниченными ресурсами**. Для каждого направления мы сообщаем данные о модели, обученной на PublicBitext, на NLLB-Seed и на их комбинации.

Экспериментальная настройка. Мы выбираем восемь направлений с ограниченными ресурсами, охваченных данными NLLB-Seed. Для справедливого сравнения с существующими наборами данных мы выбираем только те направления, для которых мы также смогли найти не менее 10 тыс. предложений общедоступного параллельного текста. Это направления ban_Latn, dik_Latn, fuv_Latn и mri_Latn, переведенные на и с английского языка (eng_Latn). Мы обучаем двуязычные модели для каждого направления, используя архитектуру трансформера с 6 слоями кодировщика и 6 слоями декодировщика, обученных с использованием расписания обучения с обратным квадратным корнем и постепенным увеличением скорости обучения. Каждая пара языков использует собственный словарь SentencePiece размером 1k. Для изучения инкрементного эффекта добавления NLLB-Seed мы подготавливаем наборы данных Diminished+NLLB-Seed для каждого из вышеупомянутых языков. Затем мы обучаем двуязычные модели и сравниваем их с PublicBitext в аналогичной настройке, как описано выше.

Результаты. Результаты в Таблице 24 показывают, что несмотря на значительно больший размер общедоступных данных для обучения, обучение на NLLB-Seed приводит к значительно более высокой производительности в среднем. Неудивительно, лучший результат достигается при комбинировании всех доступных данных. Этот результат обнадеживает, особенно учитывая последние данные, показывающие, что большую модель машинного перевода можно легко донастроить на небольшом, но высококачественном наборе данных, таком как NLLB-Seed, для адаптации к новому языку с ограниченными ресурсами (Adelani и др., 2022). Кроме того, из Таблицы 25 мы видим, что базовая производительность модели Diminished+NLLB-Seed выше, чем у базовой модели PublicBitext. Это демонстрирует, что увеличение обучающих данных модели с помощью NLLB-Seed полезно по сравнению с увеличением данных на ту же величину общедоступными данными. Для контроля за шумом мы повторяем каждый эксперимент три раза с различными подмножествами PublicBitext и сообщаем средние значения и стандартные отклонения для каждого. Небольшая вариация подтверждает, что низкие оценки на данных PublicBitext не связаны с каким-либо смещением выборки, а скорее с более низким качеством данных PublicBitext по сравнению с NLLB-Seed.

6.5.2 Эффект NLLB-Seed на обратный перевод

Мы исследуем, может ли использование NLLB-Seed дополнительно повлиять на производительность при использовании обратного перевода, например, если использование небольшого объема человечески переведенных данных, таких как NLLB-Seed, может улучшить качество обратного перевода. Для этого мы начинаем с базовых моделей, обученных на данных PublicBitext и на данных Diminished+NLLB-Seed, используем их для выполнения обратного перевода, а затем обучаем новый набор моделей на этом увеличенном наборе данных для оценки производительности.

Direction	PublicBitext chrF++	PublicBitext BT chrF++	Diminished+NLLB-Seed chrF++	Diminished+NLLB-Seed BT chrF++
Average	17.8	19.4	26.6	30.1
ban_Latn-eng_Latn	12.6 (±0.6)	14.0 (±0.3)	**22.5** (±0.2)	**24.0** (±0.1)
eng_Latn-ban_Latn	16.8 (±1.0)	18.7 (±0.3)	**22.2** (±0.2)	**25.2** (±0.3)
dik_Latn-eng_Latn	12.1 (±0.7)	13.6 (±0.2)	**16.6** (±0.2)	**17.8** (±1.0)
eng_Latn-dik_Latn	8.0 (±1.0)	10.3 (±0.2)	**13.1** (±0.4)	**14.9** (±0.1)
fuv_Latn-eng_Latn	15.6 (±0.3)	14.6 (±1.2)	**17.3** (±0.6)	**18.2** (±0.4)
eng_Latn-fuv_Latn	10.0 (±0.9)	13.6 (±0.3)	**11.8** (±0.4)	**13.1** (±0.2)
mri_Latn-eng_Latn	16.7 (±0.5)	20.6 (±0.2)	**25.9** (±0.1)	**31.8** (±0.4)
eng_Latn-mri_Latn	23.0 (±0.2)	24.8 (±0.2)	**30.6** (±0.4)	**36.2** (±0.5)

Таблица 25: **Размер обучающего набора данных и производительность валидации FLORES-200 базовых двуязычных моделей с ограниченными ресурсами и их версий, улучшенных обратным переводом**. Мы сообщаем данные о модели, обученной на общедоступных параллельных данных, и о общедоступных данных, уменьшенных и объединенных с NLLB-Seed. Каждый эксперимент повторяется три раза, и мы сообщаем средние значения с указанием стандартного отклонения в скобках.

Результаты. Из Таблицы 25 мы видим, что разрыв в производительности между моделями, обученными с использованием и без использования NLLB-Seed, дополнительно увеличивается при применении обратного перевода. Начиная только с данных PublicBitext, добавление обратного перевода дает улучшение на +1.6 по метрике chrF++, в то время как начало с данных Diminished+NLLB-Seed приводит к приросту +3.5 по метрике chrF++ (не смотря на то, что общий размер обоих наборов данных одинаков). Это указывает на

то, что небольшое количество высококачественных параллельных данных значительно улучшает эффективность методов аугментации данных на основе моделей, таких как обратный перевод.

6.6 Заключение

Повышение производительности перевода с ограниченными ресурсами в условиях массового мультиязычия сталкивается с несколькими вызовами. Прямое увеличение размера модели в значительной степени неэффективно, так как пары с ограниченными ресурсами начинают переобучаться. В этом разделе мы изучили, как наиболее эффективно увеличить емкость через смеси экспертов и представили несколько новых стратегий регуляризации. Эти методы снижают взаимное влияние между несвязанными языковыми направлениями. В сочетании с учебным планом, который вводит более ресурсные пары на более ранних этапах обучения, мы достигли значительных успехов в направлениях с ограниченными ресурсами, сохраняя при этом высокую производительность для ресурсных языков. Помимо архитектурных вызовов, повышение производительности в условиях ограниченных ресурсов затруднено из-за нехватки данных. Мы продемонстрировали, как эффективно использовать монолингвальные данные через самонадзорное обучение и более эффективную аугментацию данных. Использование данных из различных источников обратно переведенных данных от моделей массового мультиязычного перевода и традиционных систем машинного перевода, в сочетании с добытыми данными, приводит к значительному улучшению производительности.

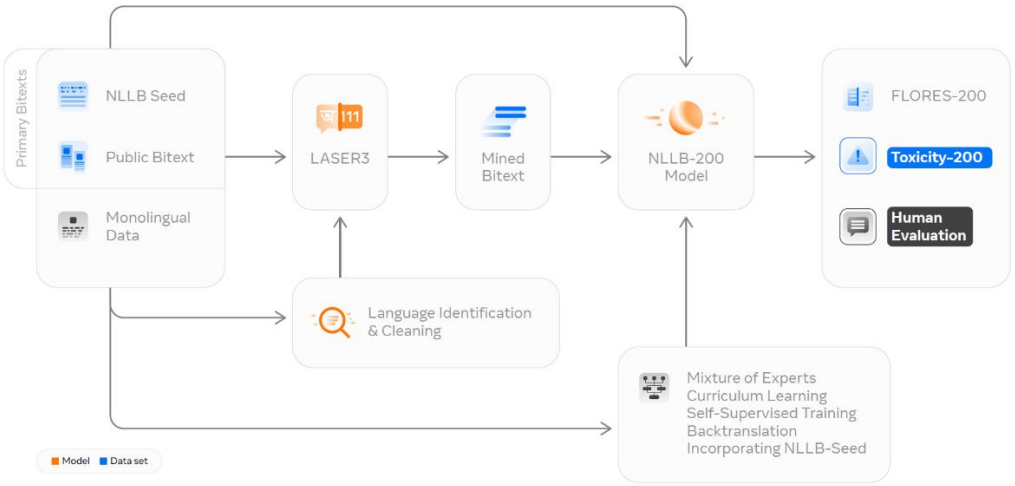


Рисунок 23: **Оценочный вклад проекта "No Language Left Behind":** Как подчеркнуто, мы описываем нашу процедуру для человеческой оценки и создания списков токсичности для более чем 200 языков.

7. Оценка

Способность количественно оценивать производительность является критической для развития систем машинного обучения, поскольку улучшение качества таких систем невозможно без надежного способа отслеживать прогресс. Машинный перевод обычно оценивается с использованием автоматических метрик, а также человеческой оценки, например, в рамках кампании оценки WMT (Akhbardeh и др., 2021), в задаче AmericasNLP Shared Task (Mager и др., 2021). В этом разделе мы описываем использованные нами автоматические метрики и методологию проведения человеческой оценки. Мы представляем результаты различных исследований, проведенных на мультиязычных моделях перевода, и анализируем надежность автоматических метрик на таком разнообразном и ресурсом-ограниченном наборе языков.

Мы также фокусируемся на анализе качества по другим аспектам. Метрики, такие как BLEU и человеческая оценка, часто фокусируются на оси качества перевода, сильно опирающейся на точность и плавность. Кроме того, мы стремимся количественно оценить наши переводы с точки зрения безопасности пользователя, выбирая акцентироваться также на количественной оценке токсичности в сгенерированных переводах. Поскольку добавление токсичного контента в общем нежелательно для пользователя, мы сначала фокусируемся на обнаружении появления такой токсичности во всех 200 языках, а затем на анализе методов смягчения и распространенности в Flores-200. Мы открываем исходный код этих новых списков токсичности для всех 200 языков.

7.1 Автоматическая оценка

Существует множество метрик для автоматической оценки качества машинного перевода, включая модельные метрики, такие как COMET (Rei и др., 2020) и BLEURT (Sellam и др., 2020). Хотя модельные метрики показали лучшую корреляцию с человеческим суждением на последних задачах метрик (Freitag и др., 2021), они требуют обучения и не могут легко расширяться на большой набор языков с ограниченными ресурсами. Другой подход - использование высоко приближенных метрик на основе круговых переводов, таких как RttLangIDChrF (Bapna и др., 2022), хотя круговые переводы могут плохо коррелировать с качеством перевода (Koehn, 2005). Такие методы легче масштабируются на новые языки, но они сильно зависят от факторов, что делает результаты сложными для воспроизведения. В данной работе мы поэтому выбираем не использовать модельные и круговые метрики, а полагаемся на BLEU и chrF++. Обе метрики опираются на основную концепцию, что качество перевода может быть количественно оценено на основе того, насколько похож перевод машины на перевод, созданный человеком. Мы кратко описываем обе метрики и их вариант BLEU ниже.

BLEU. Оценка BLEU (Papineni и др., 2002) стала стандартной метрикой оценки машинного перевода со времени её предложения два десятилетия назад. Она измеряет перекрытие между машинным переводом и человеческим референсным переводом, комбинируя точность отдельных n-грамм (от 1-граммов до 4-граммов) с штрафом за краткость.

spBLEU. Одним из крупных недостатков BLEU является его зависимость от токенизации. Усилия, направленные на стандартизацию, такие как sacrebleu (Post, 2018), добились успеха в поддержке использования токенизаторов, стандартизированных сообществом, под капотом. Однако эти токенизаторы не распространяются на многие языки. Goyal и др. (2022) предлагают spBLEU, метрику BLEU на основе стандартизированной модели SentencePiece (SPM), охватывающей 101 язык, выпущенную с Flores-101. В данной работе мы предоставляем SPM-200 вместе с Flores-200 для измерения spBLEU. Мы подробнее описываем это в разделе 8.

chrF++. Оценка chrF++ (Popović, 2017) преодолевает ограничение оценки BLEU, которая требует разделения предложения на словесные токены. Однако некоторые языки, такие как китайский или тайский, не используют пробелы для разделения слов, и инструменты сегментации слов могут отсутствовать или быть недоступными. Также существует проблема с высоко агглютинативными языками, где BLEU не присваивает кредит морфологическим вариантам. chrF++ преодолевает этот недостаток, основываясь на оценке перекрытия по n-граммам на уровне символов (n от 1 до 6) и дополняя это униграммами и биграммами слов. В данной работе мы в основном используем оценку chrF++ с настройками из sacrebleu. Однако при сравнении с другими опубликованными работами мы используем BLEU и spBLEU в случае необходимости.

7.2 Человеческая оценка

Хотя автоматические оценки являются отличным инструментом для развития исследований, человеческая оценка необходима для обеспечения осмысленной оценки качества перевода (Kocmi и др., 2021). Мы используем два современных подхода - протокол оценки XSTS и использование калибровочных наборов данных - для обеспечения осмысленных оценок качества, которые можно сравнивать между языковыми парами.

7.2.1 Методология

При создании систем машинного перевода для множества различных языковых пар ключевым вопросом является определение того, какие языковые пары достигают определенного уровня качества. Поэтому нам нужны осмысленные оценки, которые можно сравнивать между языковыми парами.

Протокол оценки XSTS. Мы адаптируем недавно предложенную методологию кросслингвистической семантической схожести текста (XSTS) от Agirre и др. (2012). Вкратце, XSTS - это протокол человеческой оценки, который фокусируется на сохранении смысла гораздо больше, чем на плавности перевода. Для языков с ограниченными ресурсами переводы обычно имеют более слабое качество, поэтому мы уделяем гораздо больше внимания использованию (сохранению смысла) переводов, даже если они не являются полностью плавными. По сравнению с прямой оценкой (Graham и др., 2013) с пятибалльной шкалой (оригинальная прямая оценка использует шкалу из 100 баллов), исследования показали, что XSTS обеспечивает более высокое согласование между аннотаторами (Licht и др., 2022).

XSTS оценивает каждое исходное предложение и его машинный перевод по пятибалльной шкале, где 1 - самый низкий балл, а 5 - самый высокий балл. Каждая точка на шкале описывается следующим образом:

1. Два предложения не эквивалентны, имеют мало общих деталей и могут касаться разных тем. Если два предложения касаются схожих тем, но менее половины основных концепций совпадают, то 1 все равно будет соответствующим баллом.

2. Два предложения имеют некоторые общие детали, но не эквивалентны. Некоторая важная информация, связанная с основной темой/глаголом/субъектом, отличается или отсутствует, что изменяет намерение или смысл предложения.

3. Два предложения в основном эквивалентны, но некоторые неважные детали могут отличаться. Не может быть значительных конфликтов в намерении или значении между предложениями, независимо от их длины.

4. Два предложения являются перефразировками друг друга. Их значения практически эквивалентны, без существенных различий или отсутствующей информации. Могут быть только незначительные различия в значении из-за различий в выражении (например, уровень формальности, стиль, акцент, потенциальные подразумеваемые значения, идиомы, общие метафоры).

5. Два предложения точно и полностью эквивалентны по значению и выражению использования (например, уровень формальности, стиль, акцент, потенциальные подразумеваемые значения, идиомы, общие метафоры).

Калибровочный набор. Для обеспечения осмысленных оценок, которые можно сравнивать между языковыми парами, мы просим каждого оценщика предоставить оценки с использованием шкалы XSTS на одинаковом наборе предложений. Это делается для идентификации того, какие группы аннотаторов имеют системную тенденцию быть более строгими или щедрыми в оценках, и корректировки этого эффекта. В то время как оценщики оценивают различные языки, калибровочный набор состоит из машинных переводов на английский язык в паре с английским референсным переводом. На основании того, как оценщики используют шкалу XSTS на этом калибровочном наборе, мы корректируем их исходные оценки на фактической оценочной задаче для обеспечения последовательности оценок между оценщиками. Хотя монолингвистическая задача не совсем точно имитирует билингвальную задачу XSTS, это разумное первое приближение и показало увеличение корреляции между человеческими и автоматическими метриками, в основном за счет снижения одного источника «шума» в человеческих оценках - отсутствия калибровки оценок между аннотаторами.

Получение совокупных метрик качества по оценкам человека из нескольких исследований. Для получения совокупной метрики качества по оценкам человека для каждой группы предложений, оцененных теми же аннотаторами, мы выбираем наиболее часто

встречающуюся оценку XSTS для каждого предложения и усредняем эти наиболее часто встречающиеся оценки по всем оцениваемым предложениям. В рамках данного исследования совокупная оценка человеческой оценки для пары исходного и целевого языков ls -> lt

$$H_{l_s \to l_t} = \frac{1}{|\mathcal{T}_{l_s \to l_t}|} \sum_{(S,T) \in \mathcal{T}_{l_s \to l_t}} \text{median}\{X_{l_s \to l_t, i}(S,T) \mid 1 \leq i \leq M_{l_s \to l_t}\}, \qquad (14)$$

где ls -> lt обозначает пару исходного и целевого языков, Xls!lt;i(S; T) обозначает оценку XSTS i-го оценщика, оценивающего предложения в заданном направлении перевода ls -> lt для исходного предложения S и целевого предложения T. Mls->lt обозначает общее количество оценщиков, оценивающих направление перевода ls -> lt, и Tls->lt = f(Sls->lt;k; Tls->lt;k) j 1 k Nls->ltg - это набор Nls!lt (исходное, целевое) пар предложений, оцениваемых для направления перевода ls ! lt.

Каждый оценщик в данном исследовании также запрашивается оценивать калибровочный набор Cs = f(Ss;k; Ts;k) j 1 k Ksg, где Ss;k обозначает k-е исходное предложение в калибровочном наборе для исследования s, Ts;k обозначает целевое предложение, соответствующее Ss;k, и Ks = jCsj - количество предложений в калибровочном наборе для исследования оценки s. Оба исследования имели одинаковые калибровочные наборы Cs = C, состоящие из K = Ks = 1000 обратно переведенных предложений Flores-200 различного качества. Для каждого неанглийского языка, оцененного в исследовании, мы получаем среднюю медианную оценку XSTS на калибровочном наборе:

$$C_l^{(s)} = \frac{1}{|\mathcal{C}|} \sum_{(S,T) \in \mathcal{C}} \text{median}\{X_{l,i}^{(s)}(S,T) \mid 1 \leq i \leq M_l\}, \qquad (15)$$

где l обозначает неанглийский язык, оцениваемый в исследовании, X(s)l;i (S; T) обозначает оценку XSTS, предоставленную i-м оценщиком для неанглийского языка l в исследовании s для исходного предложения S и переведенного предложения T. Обратите внимание, что в пределах одного исследования набор оценщиков для некоторого направления перевода ls ! lt совпадает с набором оценщиков для обратного направления перевода lt ! ls, если обе пары оцениваются в рамках исследования. Следовательно, есть только одно калибровочное значение для каждого языка.

Чтобы получить откалиброванные совокупные оценки XSTS для группы переводов в заданном исследовании s, мы вычитаем скорректированную среднюю медианную оценку XSTS на калибровочном наборе из средней медианной оценки XSTS на оценочном наборе:

$$\widetilde{H}_{l_s \to l_t}^{(s)} = H_{l_s \to l_t}^{(s)} - (C_{L(l_s \to l_t)}^{(s)} - \bar{C}), \qquad (16)$$

где L(ls ! lt) обозначает неанглийскую часть языковой пары ls ! lt, а C - средняя калибровочная оценка по всем неанглийским языкам l, оцененным во всех исследованиях, т.е.

$$\bar{C} = \frac{\sum_s \sum_{l \in \mathcal{L}(s)} C_l^{(s)}}{\sum_s \sum_{l \in \mathcal{L}(s)} 1}, \qquad (17)$$

с L(s) = {L(ls ! lt) | ls ! lt оценен в s} обозначается множество неанглийских языков, оцененных в исследовании s.

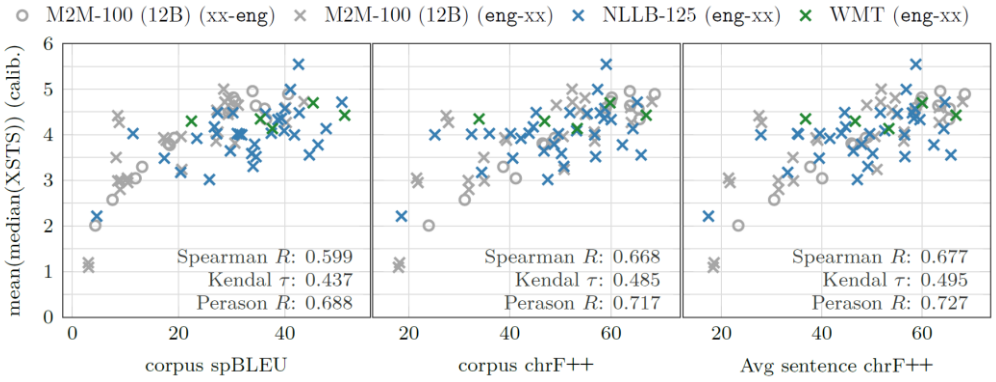


Рисунок 24: **Корреляции между агрегированными оценками качества от людей и автоматическими метриками**. Левая часть показывает связь между spBLEU и XSTS, средняя часть - между chrF++ и XSTS, а правая часть - между средним chrF++ на уровне предложения и XSTS. Все автоматические оценки вычислялись только для предложений, оцененных для данной модели и направления перевода (либо полного набора данных Flores-200, либо его подмножества).

7.2.2 Результаты

Производительность моделей машинного перевода, оцененная человеческими экспертами, была подробно проанализирована для двуязычных моделей и специфических областей. Например, ежегодные оценки на семинаре по машинному переводу (Akhbardeh и др., 2021) рассматривают несколько направлений перевода в новостной области. Другая известная кампания оценки (IWSLT) акцентирует внимание на переводе речи (Anastasopoulos и др., 2021). В нашем же исследовании мы сосредотачиваемся на многоязычном переводе. В этом разделе мы анализируем корреляцию между оценками качества от людей и автоматическими метриками, такими как chrF++, изучаем сложность набора данных Flores-200 с точки зрения человеческих оценщиков в предварительных исследованиях и обсуждаем вариации оценок качества перевода от людей в различных языковых средах.

Исследования качества перевода по оценке людьми. Хотя человеческая оценка является золотым стандартом для понимания истинного качества перевода, автоматическая оценка критична для проектирования моделей. Например, сравнение производительности 10 моделей при параметрическом поиске будет зависеть от автоматических метрик. Мы использовали агрегированные результаты двух масштабных многоязычных исследований оценки качества от людей для изучения взаимосвязи между человеческими мерами качества и автоматическими оценками, такими как spBLEU и chrF++. Эти исследования включают оценку переводов трех различных моделей перевода (M2M-100 12B (Fan и др., 2020), NLLB-125 - модель MoE, охватывающая 125 языков, и многоязычное англоцентричное представление WMT2021, охватывающее 7 языков (Tran и др., 2021)) и 49 различных направлений перевода, оцененных до 157 отдельными человеческими оценщиками. Для каждого масштабного исследования оценки каждой комбинации модели перевода и направления перевода были назначены 3 оценщика, которые оценивали качество переводов на подмножестве набора данных Flores-200. Из 79 уникальных комбинаций исследований оценки, направлений перевода и моделей перевода, которые были оценены в двух исследованиях, 60 были оценены на полном наборе данных Flores-200, а 19 были оценены на половине набора данных Flores-200.

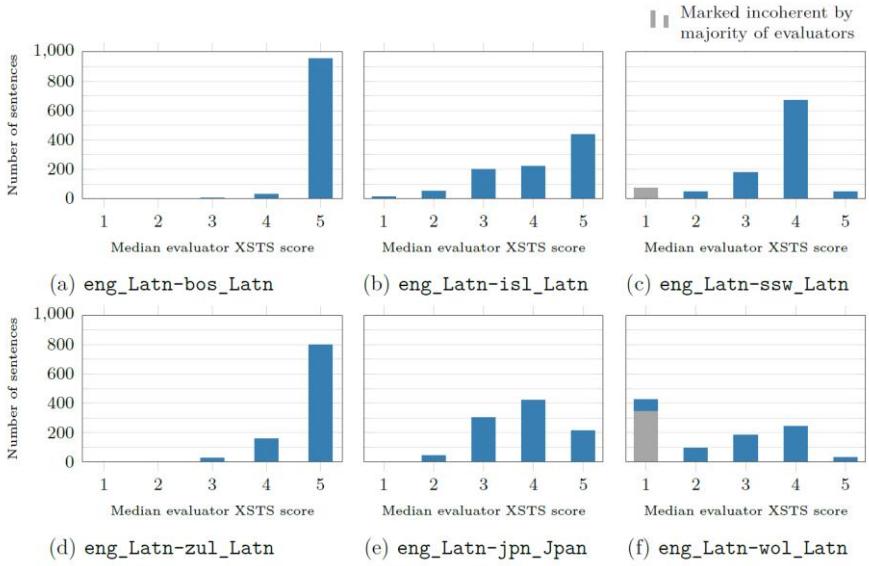


Рисунок 25: **Выбранные распределения медианных оценок XSTS для различных направлений перевода**. Все переводы были сгенерированы моделью NLLB-125, и все оценки взяты из одного исследования оценки. Показаны три различных распределения.

Как коррелирует человеческая оценка с chrF++ и spBLEU? Мы обнаружили, что автоматизированные метрики, такие как spBLEU и chrF++, разумно хорошо коррелируют с откалиброванными человеческими оценками качества перевода, как показано на рисунке 24. В частности, коэффициенты корреляции Спирмена R между агрегированными XSTS и spBLEU, chrF++ (корпус) и chrF++ (средний уровень предложения) составляют 0,599, 0,668 и 0,677 соответственно. Другие коэффициенты корреляции (Кендалла и Пирсона) имеют тот же порядок. Усреднение chrF++ на уровне предложения дает лучшую корреляцию с нашим агрегированным XSTS-счетом, что вероятно является результатом выбора нашего метода агрегации: XSTS представляет собой среднее значение медианной оценки для всех оцениваемых предложений, независимо от их длины.

Мы также обнаружили, что откалиброванные оценки человеческой оценки коррелируют с автоматизированными оценками сильнее, чем неоткалиброванные оценки человеческой оценки по всем автоматизированным метрикам и выборам коэффициента корреляции. В частности, неоткалиброванные оценки человеческой оценки имеют коэффициент корреляции Спирмена R равный 0,557, 0,586 и 0,592 для spBLEU, chrF++ (корпус) и chrF++ (средний уровень предложения) соответственно.

Как различаются оценки человеческой оценки по разным языкам? Мы также рассматриваем индивидуальные распределения оценок для модели NLLB-125. Мы наблюдаем три грубые категории распределения оценок XSTS. Первая категория - высокая производительность в целом, что означает, что все предложения для этого направления перевода оценены высоко.

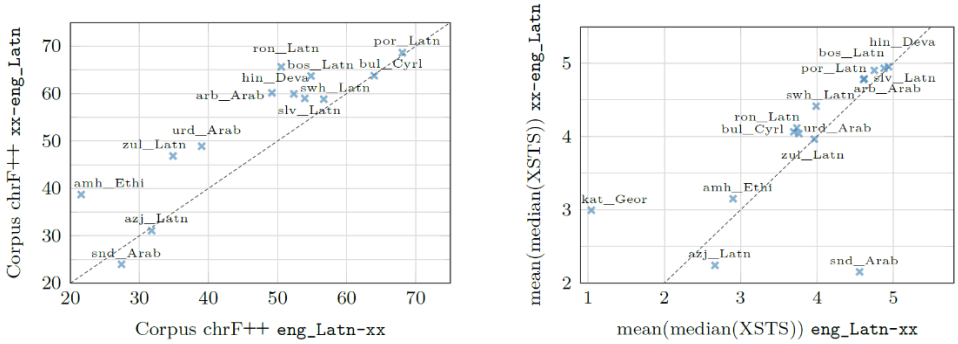


Рисунок 26: **Сравнение качества перевода на английский и с английского**. На левом графике сравниваются оценки chrF++, а на правом - средние медианные оценки XSTS для направлений перевода на английский и с английского языки. Все переводы выполнены с использованием модели перевода M2M-100 (12B).

Это отображается в первом столбце Рисунка 25. Такое распределение указывает на то, что почти все оцененные предложения показывают высокую производительность. Второй шаблон показан во втором столбце Рисунка 25, демонстрируя разнообразное качество. Для этих языков средняя оценка может быть высокой, но есть много предложений, которые оцениваются как плохие. Наконец, третий шаблон, который мы наблюдаем, включает большое количество низкокачественных переводов (оценка XSTS 1) вместе с высоким процентом несвязных предложений, что означает, что оценщик отметил перевод как несвязный. Эти данные показаны в третьем столбце Рисунка 25 и часто представляют собой текст, который в основном непонятен или имеет полностью искаженное смысловое содержание.

Оценка качества перевода на английский и с английского языки. Несколько предыдущих работ (Arivazhagan et al., 2019) и наши результаты в разделе 6 показывают, что перевод с различных языков на английский язык дает более высокие оценки BLEU, чем перевод с английского. Мы сравниваем различия в человеческой оценке качества перевода на английский и с английского языки в Рисунке 26. В целом мы находим, что, как подтверждают автоматические оценки, такие как chrF++ и spBLEU, оценки качества перевода, полученные от людей, также отражают, что качество перевода на английский язык обычно лучше, чем качество перевода с английского языка, за исключением некоторых случаев, таких как snd и azj, где качество перевода на английский язык значительно хуже, как по автоматическим, так и по человеческим метрикам оценки.

7.3 Токсичность

Обнаружение токсичности в цифровом контенте получило значительное внимание в последние годы как для пользовательского языка (Kiritchenko et al., 2021; Mishra et al., 2019; Vidgen et al., 2019; Zampieri et al., 2019), так и для машинно-генерированного текста (Bender et al., 2021; Xu et al., 2020). Генерация токсичного контента исследовалась в различных задачах классификации предложений и диалогов, но не в достаточной мере в области перевода. Однако добавление токсичных терминов, то есть перевод контента, содержащего токсичные слова, которых не было в исходном тексте, может иметь значительно негативное влияние на общее качество перевода. Это, например, может привести к утрате доверия пользователей. Наша цель в данном разделе - провести анализ токсичности в мультиязычных моделях машинного перевода. Мы предоставляем первоначальную базовую оценку токсичности для оценки в большом количестве языков путем сбора и публикации списков токсичных слов в 200 языках (Раздел 7.3.1). Далее, в Разделе 7.3.3 мы предлагаем и оцениваем простые, но масштабируемые детекторы токсичности, которые могут быть оптимизированы по точности или полноте в зависимости от конкретного применения (например, фильтрация или обнаружение). Затем мы предлагаем стратегию фильтрации для снижения дисбаланса токсичности в обучающих данных и визуализируем вклад источников нескольких примеров с добавленной токсичностью в Разделе 7.3.3. Наконец, мы обсуждаем открытые проблемы и этические аспекты в Разделе 7.3.4. Обратите внимание, что в этом разделе мы будем приводить примеры переводов, содержащих токсичный язык.

7.3.1 Введение

Что такое токсичность? Токсичность в обработке естественного языка может быть определена как использование слов или фраз, вызывающих оскорбительные высказывания и негативные настроения (Google Jigsaw, 2017). В контексте перевода токсичность может изначально присутствовать в исходном тексте или быть созданной в процессе перевода (добавленная токсичность). Эта добавленная токсичность может возникать из-за неправильного перевода (например, неверного выбора лексики) или как результат генерации нового целевого слова "с нуля", что ведет к неточным переводам.

Цель обнаружения и смягчения токсичности. Надежные универсальные системы машинного перевода должны способны адекватно переводить любое исходное содержимое независимо от области или стиля регистрации, что включает перевод языка, который может рассматриваться как токсичный. Однако они должны оставаться верными исходному содержанию и не добавлять в процессе перевода элементы токсичности, которых нет в исходном тексте. Наша основная цель - повышение безопасности перевода путем минимизации вероятности катастрофических неправильных переводов. Следует отметить, что добавленная токсичность представляет собой один из видов катастрофических неправильных переводов (Specia et al., 2021), наряду с неправильными переводами именованных сущностей, рода (Levy et al., 2021), чисел и единиц измерения, а также изменениями семантической полярности.

7.3.2 Списки токсичности для 200 языков

Для возможности обнаружения токсичности в масштабе мы сосредотачиваемся на подходе на основе списков слов. В этом разделе мы описываем, что мы считаем токсичным элементом для включения в список и как мы масштабируем создание этих списков на сотни языков.

Какой контент считается токсичным? Из-за субъективной природы токсичности определения токсичного языка могут варьироваться. Мы включаем элементы, которые обычно считаются вульгарными или непристойными (Google Jigsaw, 2017). В дополнение к ним, мы включаем элементы, более специфически связанные с изображениями порнографического содержания или сексуальными актами, некоторые часто используемые выражения ненависти и некоторые формы издевательств (включая язык, который может вызывать травму или быть использованный с целью заткнуть кого-то). Мы также включаем элементы, вульгарные или нет, относящиеся к частям тела, которые обычно ассоциируются с сексуальными практиками.

Методология сбора. Токсичность является культурно чувствительной, что представляет собой вызов при начале работы с одного исходного языка (в данном случае, английского) и попытке найти эквиваленты в таком масштабном мультиязычном окружении. Термины ненависти, такие как расистские или этнические оскорбления, например, могут оказаться самыми сложными. Мы начинаем с профессионального перевода первоначального списка, составленного на английском языке, и допускаем дополнения для адаптации к культурным особенностям.

Мы итеративно разработали шаблон для перевода токсичности, который предоставляет информацию для целей разъяснения и контекстуализации. В последней версии шаблона дополнительная информация включает разделение по доменам (например, оскорбления, сексуальные термины, сокращения), информацию о части речи, ссылки на словарные определения слов в их токсичном смысле, указания на языковые регистры (жаргон, вульгарный, официальный и т. д.).

Кроме того, шаблон предусматривает четко определенные места для добавления морфологических вариантов (если целевой язык богат морфологически). Для полисемичных терминов, которые могут быть токсичными в зависимости от контекста, шаблон предлагает дополнительное пространство и рекомендации относительно лучших практик разъяснения через предложение менее неоднозначных коротких n-граммов (обычно, $0 < n < 4$). Для уменьшения культурных слепых зон другой раздел шаблона дает переводчикам возможность вставлять общие токсичные выражения, для которых может быть сложно найти прямой английский эквивалент. Переводчиков просят дать объяснения или дословные описания. Предложения ограничены примерно сорока элементами без конкретных ограничений на количество производных словоформ на каждый элемент, за исключением общего руководства о необходимости оставаться в рамках часто используемых словоформ (то есть избегать редких и архаичных словоформ). Переводчиков не просят создавать литспик или нестандартные орфографические варианты, но не запрещают включать их, если это сочтется целесообразным. Наконец, шаблон позволяет второму переводчику действовать как рецензент и вносить комментарии и дополнительные предложения.

Результатом является список n-грамм, представляющих в основном: общие ругательства, часто используемые оскорбления, порнографические термины, часто используемые выражения ненависти, некоторые термины, используемые для издевательств, и части тела, обычно ассоциируемые с сексуальной активностью.

Краткий обзор списков токсичности. Для подведения итогов наших списков обнаружения токсичности по всем 200 языкам средняя длина списка составляет 271 запись, а медианное число записей - 143. Последнее может быть более точной мерой центральной тенденции по сравнению с средним значением, учитывая, что языки с богатой флективной морфологией составляют крайние выбросы (например, чешский список имеет 2,534 записи, польский список - 2,004). Самый короткий список содержит 36 записей, а самый длинный - 6,078.

Связанные работы. Для обнаружения токсичности в естественном языке можно использовать различные подходы, основанные на списочных методах или на методах машинного обучения. Большинство недавних анализов токсичности в NLP основаны на обучении надзорных классификаторов на уровне предложений в основном для английского языка и расширяются до нескольких других языков (Lees et al., 2021) в многоязычной классификации. Однако нам не известен анализ машинного перевода, который оценивает добавленную токсичность или дисбаланс токсичности в параллельных тренировочных наборах данных и масштабирует эти анализы на большое количество языков. В целом, для обучения надзорных классификаторов требуется большое количество размеченных данных, что часто ограничивает их использование высокоресурсными языками.

Однако, что более важно, недавние исследования утверждают о сильных предвзятостях этих классификаторов в неправильном ассоциировании языка представителей недостаточно представленных социальных групп с токсичностью (Xu et al., 2021a). Хотя мы не можем обнаружить нетекстовую токсичность или токсичные термины, которые не включены в наши списки слов, наш подход более масштабируем и, следовательно, способен расширяться на сотни языков. Кроме того, использование подхода на основе списков слов обеспечивает более высокую степень прозрачности и объяснимости. Наконец, наши списки слов могут потенциально использоваться для других приложений в области NLP. Возможные ограничения списков слов обсуждаются в разделе 7.3.4.

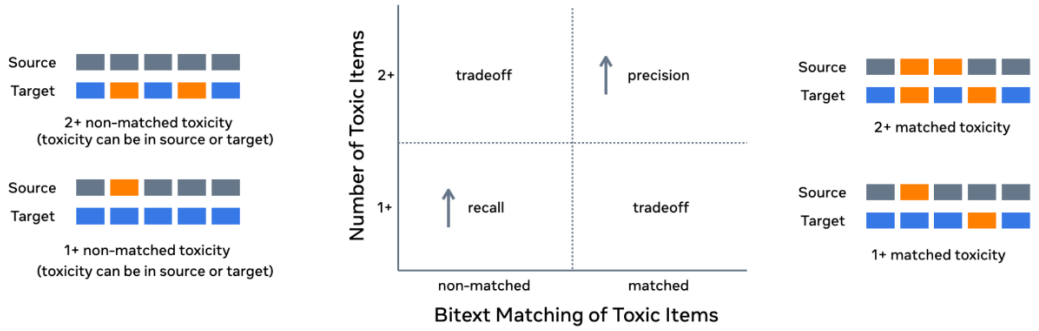


Рисунок 27: **Обзор детекторов токсичности**, варьирующихся по числу токсичных элементов и соответствию битекста токсичным элементам.

7.3.3 Обнаружение токсичности

Методология. Наши детекторы токсичности определяют токсичность с использованием двух критериев: количество токсичных элементов и сопоставленная токсичность. Это иллюстрируется на Рисунке 27. Мы определяем токсичные элементы как короткие n-граммы (обычно с $0 < n < 4$), присутствующие в наших списках, описанных в разделе 7.3.2. Среди числа токсичных элементов мы рассматриваем два случая: 1 или более токсичный элемент и 2 или более токсичных элемента. Токсичность может быть сопоставленной или несопоставленной, где несопоставленная означает, что токсичные элементы присутствуют либо в исходной, либо в целевой части битекста. В отличие от этого, сопоставленная токсичность означает, что токсичные элементы присутствуют с обеих сторон битекста, но это не обязательно означает правильный перевод токсичных слов.

Поэтому у нас есть следующие варианты детекторов токсичности: (1) 1+ токсичный элемент несопоставленный, (2) 2+ токсичных элемента несопоставленных; (3) 1+ токсичный элемент сопоставленный; (4) 2+ токсичных элемента сопоставленных.

Токсичный элемент (или фраза) считается обнаруженным, если он присутствует в строке для анализа и окружен пробелами или началом/концом строки (например, мы не обнаруживаем "bass", если ищем "ass"). Мы отслеживаем количество уникальных токсичных элементов, найденных в строке, но не учитываем фразу повторно, если она встречается несколько раз. Для некоторых языков (например, ассамский, бирманский, ория, корейский, китайский) токенизация по пробелам недостаточна для различения слов друг от друга. В таких случаях мы используем токенизацию SentencePiece как для предложений, так и для токсичного словника.

Оценка. Для количественной оценки качества наших списков токсичности и детекторов мы планируем использовать общепринятый бенчмарк токсичности перед их применением к машинному переводу. Мы используем набор данных Jigsaw,33 который направлен на анализ входного предложения и определение, является ли это предложение токсичным или нет. Этот набор данных классифицирует токсичность среди следующих подклассов: токсичный, любой токсичный, неприличный, угроза, оскорбление, сильно токсичный, ненависть к личности.

Мы использовали тестовый набор данных, доступный из первого челленджа, и отфильтровали комментарии, применяя следующие критерии: строки с менее чем 600 символами, менее чем 100 словами и с хотя бы 1 словом-токеном. Затем мы отфильтровали только те строки, которые фактически содержали хотя бы 1 английскую букву, чтобы исключить только цифры, эмодзи и т. д. Наш окончательный набор данных составляет 86.2% безобидных предложений, а оставшиеся 13.8% предложений имеют 1 или более меток токсичности (с неисключительными классами).

Поскольку нет сравнимой базовой линии по такому большому числу языков, мы предоставляем производительность лучше случайной модели базовой линии, которую вычисляем как случайную генерацию меток токсичности/не токсичности (как на случайных 50/50 токсичности/не токсичности, так и на 13.8%/86.2% токсичности/не токсичности). Затем мы используем модель перевода NLLB-200 для перевода этого английского набора во все доступные языки в модели. Следует отметить, что использование модели перевода для генерации неанглийских ссылок на токсичность повлияет на анализ качества наших детекторов, так как перевод может изменить уровень токсичности в ссылках. Однако у нас не было альтернативы для оценки данных с метками токсичности для наших 200 языков.

Таблица 26 отображает результаты наших детекторов для английского, французского (как пример языка с высокими ресурсами) и среднее значение по всем языкам с chrF++ > 45. Лучший детектор токсичности показывает F1-оценку 66.8% для английского и 61.6% для высокоресурсного языка французского. Эти показатели значительно превосходят наши базовые линии лучше случайного выбора. На Рисунке 28 показано снижение F1-оценок для детекторов с 1+ токсичным элементом по сравнению с детекторами с 2+ токсичными элементами. Пунктирные линии на графике представляют производительность случайного выбора. Набор языков представляет собой равномерную выборку нескольких языков с высокими и низкими ресурсами. Для минимизации смещающего фактора качества перевода и детекторов мы ограничили выборку теми, у кого chrF++ выше 45, как показатель качества модели.

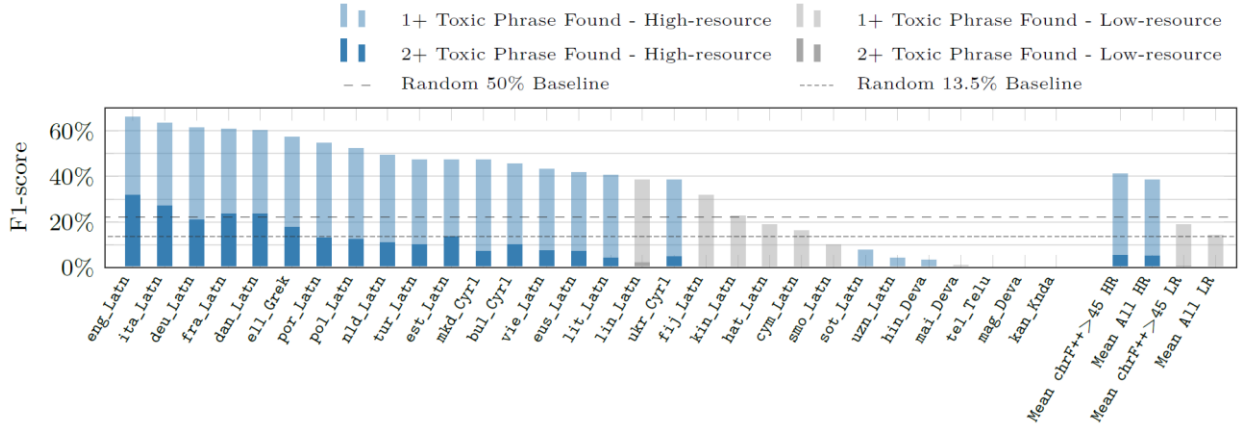


Рисунок 28: **Упорядоченные значения F1 для детекторов токсичности на основе списков**. 1+ токсичный элемент против 2+ токсичных элементов. Пунктирные линии показывают производительность случайных базовых уровней. Набор языков представляет собой равномерную выборку нескольких языков с высоким и низким уровнем ресурсов, ограниченную теми, у кого chrF++ > 45. Представлены средние оценки как для всех языков, так и для языков с учетом порога chrF++.

Среднее значение вычисляется как для языков с chrF++ выше 45, так и для всех языков, без учета порога chrF++. В среднем наши детекторы работают лучше случайных ожиданий, даже для языков с ограниченными ресурсами, что демонстрирует полезность наших детекторов. Однако существует значительная разница между языками с высоким и низким уровнем ресурсов. Результаты для языков с наименьшей производительностью на Рисунке 28, таких как хинди (hin_Deva), каннада (kan_Knda), майтили (mai_Deva), телугу

(tel_Telu) и магахи (mag_Deva), могут частично объясняться тем, что скрипты, на которых написаны эти языки, не всегда адекватно токенизированы нашими детекторами.

Оценка токсичности в обучающих **данных**. Мы использовали наши детекторы (1+ токсичный элемент несоответствующий) для подсчета потенциально токсичных элементов в данных обучения битекстов. На Рисунке 29 показан процент токсичных элементов в корпусах только с английской стороны наших пар битекстов. Рисунок 30 показывает процент токсичных элементов в корпусах с не-английской стороны. Поскольку мы сравниваем данные по разным языкам, мы используем корпус Христианской Библии для калибровки базового уровня. Мы используем этот корпус, поскольку он является многоязычным и должен быть наиболее последовательным по содержанию, качеству и уровню токсичности среди языков. Если бы корпуса Христианской Библии не было доступно, результаты остались бы некалиброванными.

Мы видим, что уровень потенциальной токсичности варьируется среди языков и корпусов. В целом, OpenSubtitles и qed имеют большее количество потенциально токсичных элементов в нескольких языках, а Mined Data имеет меньшее количество. Мы изучили выборку потенциальной токсичности этих корпусов и обнаружили, что значительная часть из них представляет собой неправильно выровненные битексты. Токсичность присутствует только на одной стороне этих битекстов, либо через полное отсутствие (что является наиболее распространенным случаем), либо через значительную детоксикацию. Обучение на таких неправильно выровненных битекстах может способствовать ошибочным переводам с добавлением токсичности. Смягчение токсичности путем фильтрации обучающих данных. Дополнительная токсичность может возникнуть из токсичных элементов, присутствующих в обучающих данных. Цель фильтрации данных состоит в выявлении дисбалансов в токсичности, которые могут в конечном итоге привести к созданию токсичных неправильных переводов.

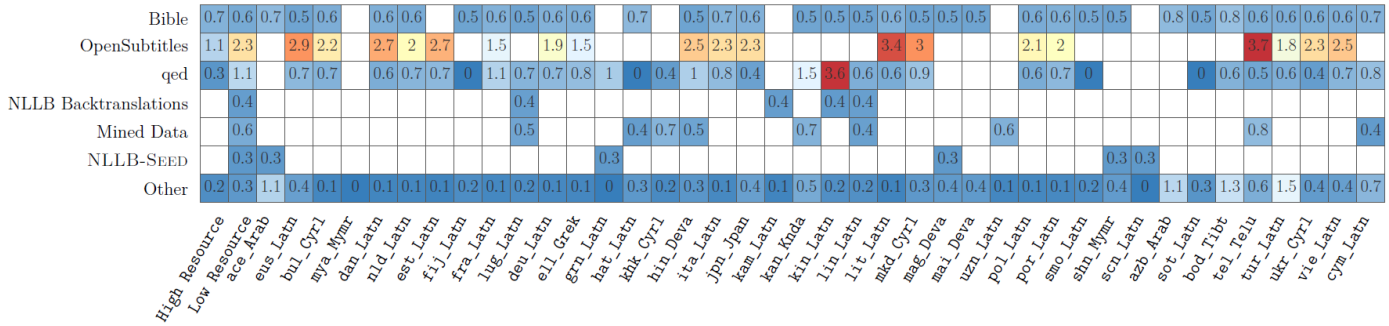


Рисунок 29: **Процент токсичных элементов, найденных в каждом корпусе по английской стороне**. Один и тот же список использовался для всех пар языков в корпусах.

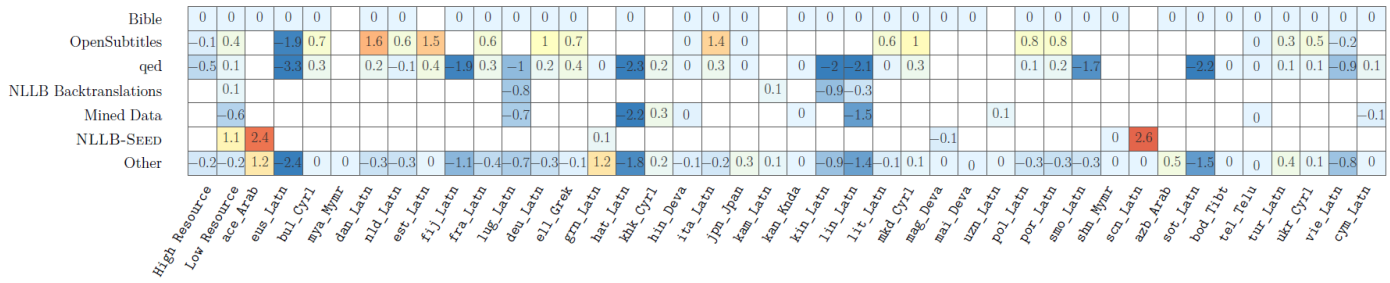


Рисунок 30: **Процент токсичных элементов, найденных в каждом корпусе на не-английской стороне**, откалиброванный путем вычитания значения корпуса Библии, для обеспечения согласованности между языками.

Имейте в виду, что мы не хотим фильтровать всю токсичность из обучающих корпусов, потому что это может привести к искажению и помешать модели правильно переводить потенциально токсичные элементы, даже в безобидном контексте. Таким образом, нас интересует удаление предложений, действительно содержащих токсичные элементы только с одной стороны, при оптимизации для точности. Анализ качества детектора показывает, что когда наши детекторы обнаруживают несколько случаев токсичности (2+ токсичных элемента) в предложении, мы получаем более высокую точность в обнаружении токсичности. На основе этого мы предлагаем фильтровать обучающие предложения, содержащие различия в нескольких токсичных элементах при сравнении исходной и целевой сторон.

Для проверки этой методологии мы включили фильтрацию токсичности в процесс фильтрации, который будет подробно описан в разделе 8.1.4. Затем мы обучили двуязычные модели с использованием и без использования этой фильтрации с использованием архитектуры, описанной в разделе 6.5. Таблица 27 показывает результаты на наборе данных Flores-200 dev. В таблице указаны количество обучающих предложений, качество перевода (chrF++) и количество токсичных элементов в выходных данных перевода. Все цифры приведены до и после фильтрации. Мы наблюдаем, что процесс фильтрации, включающий фильтрацию токсичности, не только улучшает производительность перевода, но также обычно снижает количество токсичных элементов в переводе.

Интерпретируемость добавленной токсичности. Добавленная токсичность в выходных данных перевода может возникать из галлюцинации или ошибки перевода. Следует отметить, что галлюцинированная токсичность означает, что в целевом выходе есть токсичный элемент, который появляется без прямого соответствия в исходном тексте. В отличие от этого, ошибочно переведенная токсичность означает, что мы переводим не токсичное исходное слово в токсичное целевое слово. Для различения этих двух случаев

мы используем ALTI+ (Ferrando и др., 2022), который позволяет исследовать вклады исходных и целевых префиксов в прогнозы модели. Рисунок 31 отражает входные атрибуции определенного слова.

	Before Filtering			After Filtering		
Direction	# Train	chrF++	Toxicity	# Train	chrF++	Toxicity
eng_Latn-smo_Latn	7.3 M	48.7	33	3.8 M	**49.4**	**22**
eng_Latn-sot_Latn	10.2 M	44.4	9	6.2 M	**44.4**	**6**
eng_Latn-twi_Latn	7.1 M	35.7	138	2.9 M	**36.7**	**136**
eng_Latn-umb_Latn	1.2 M	13.1	1	0.8 M	**23.3**	**0**
eng_Latn-vec_Latn	8.1 M	37.7	**14**	3.8 M	**40.6**	22
smo_Latn-eng_Latn	43.0 M	49.0	4	32.2 M	**49.7**	**4**
sot_Latn-eng_Latn	42.4 M	51.8	2	33.4 M	**52.3**	**2**
twi_Latn-eng_Latn	41.5 M	38.4	1	30.1 M	**39.8**	**1**
vec_Latn-eng_Latn	41.4 M	53.8	2	30.1 M	**56.3**	**2**
umb_Latn-eng_Latn	39.3 M	27.1	4	29.9 M	**27.2**	**2**

Таблица 27: **Эксперименты с фильтрацией** для двуязычных моделей перевода. Мы сообщаем для каждого направления количество обучающих предложений, значение chrF++ и количество добавленных токсичных элементов.

Каждый пример содержит исходное предложение, предложение целевого вывода и предложение эталонного перевода на английский язык.

• Предложение 1, которое переводится с волофа на английский, показывает вклад входных данных слова "idiots". Мы видим, что вклад исходного слова "idiots" сосредоточен на целевом префиксе слова "self-absorbed". Общий вклад исходного слова для предсказанного слова "idiots" составляет 0.35, что значительно меньше общего вклада целевого префикса, равного 0.65. Опять же, это согласуется с тем, что это слово **галлюцинировано**, так как оно не имеет прямого соответствия исходному токену.

• Предложение 2, которое переводится с кечуа на английский, показывает вклад входных данных слова "ass". Мы видим, что вклад исходного слова "ass" сосредоточен на целевых словах "pain, in, the". Общий вклад исходного слова для предсказанного слова "ass" составляет 0.20, что значительно меньше общего вклада целевого префикса, равного 0.80. Это согласуется с тем, что это слово **галлюцинировано**, так как оно не имеет прямого соответствия исходному токену.

• Предложение 3 показывает вклад входных данных слова "penis", при переводе с северного курдского на английский. Мы видим, что эти вклады ясно включают исходное слово "Penceya". Общий вклад исходного слова для предсказанного слова "penis" составляет 0.68, что значительно выше общего вклада целевого префикса, равного 0.32. Мы приходим к выводу, что это слово **переведено** неверно, так как оно имеет прямое соответствие исходному токену "Penceya", которое должно быть переведено как "коготь".

• Наконец, предложение 4 показывает вклад входных данных слова "Nazis", при переводе с немецкого на английский. Общий вклад исходного слова для предсказанного слова "Nazis" составляет 0.77, что значительно выше общего вклада целевого префикса, равного 0.23. Этот пример добавлен для сравнения поведения ALTI+ в **точном** токсичном переводе.

Hallucinated Toxicity	**Mistranslated Toxicity**
Sentence 1	Sentence 3
Montaru biologie ay osiyatër yu yor seen bopp lañu yuy wey ci ab diir bu moom boppam donte amul ay juñju biti.	Penceya wî ya didoyê mezintir bû, û bû sebebê navê Hesperonychus-ê
It's about a bunch of self-absorbed **idiots** who live in their own little world, and they don't have time for anything else.	His dodo's **penis** was larger, and was the reason for the name Hesperonychus
Reference 1	Reference 3
Biological clocks are self sustaining oscillators which will continue a period of free-running cycling even in the absence of external cues.	Its second claw was larger, giving rise to the name Hesperonychus which means "western claw."
	Correctly Translated Toxicity
Sentence 2	Sentence 4
Kayqa nanaykuna tañichiqina imaymanamkunapi.	Machen Sie keine Witze über den Holocaust oder die Nazis.
It's a pain in the **ass.**	do not make jokes about the Holocaust or the **Nazis.**
Reference 2	Reference 4
This is just like symptomatic treatment in many cases.	Do not make jokes about the Holocaust or **Nazis**.

Рисунок 31: **Визуализации входных вкладов источника и цели**. Галлюцинированная токсичность (предложение 1) показывает входные вклады для токсичного элемента "idiots" при переводе с волофа на английский; (предложение 2) показывает входные вклады для токсичного элемента "ass" при переводе с кечуа на английский; Переведенная неправильно токсичность (предложение 3) показывает входные вклады для токсичного элемента "penis" при переводе с курдского на английский. В каждом примере показана английская ссылка; Правильно переведенная токсичность (предложение 4) показывает идеальный перевод для сравнения с предыдущими примерами.

7.3.4 Открытые вызовы в области токсичности для машинного перевода

Существует множество открытых исследовательских вопросов по вызову обнаружения токсичности в масштабе для сотен языков. Поскольку мы оцениваем наш подход на переведенном наборе данных, качество перевода может стать фактором, влияющим на исследование. Например, качество обнаружения токсичности может зависеть от доступных ресурсов для каждого языка. Кроме того, качество и эффективность наших детекторов, которые обнаруживают или фильтруют токсичность, может изменяться в зависимости от длины списка слов, точности сегментации, вариации класса токсичности и количества нелексикализированной токсичности. Среди вызовов в списке слов включены морфологические аспекты, такие как чувствительность к регистру. Различия в регистрах часто связаны с порядком слов; например, первые символы слов в начальной позиции предложения часто пишутся с заглавной буквы. Наши детекторы приводят все элементы к нижнему регистру перед обнаружением, что делает невозможным различение некоторых омографов, отличающихся только регистром первых символов. Расширение и разрешение малых списков токсичности являются критическими областями для будущей работы, которая, вероятно, потребует тесного сотрудничества с большим количеством носителей языка. Первым шагом к дисамбигуации может быть контекстуализация полисемичных слов путем замены одиночных токенов n-граммами, которые намного вероятнее представляют настоящее токсичное содержимое. Наконец, мы знаем, что добавленная токсичность может быть вызвана явлением галлюцинации или ошибкой в переводе. Наши примеры визуализации с помощью ALTI+ показывают, что низкий вклад источника в токсичный элемент, вычисленный с использованием этого метода, является сильным индикатором галлюцинации. Мы хотим использовать эту информацию для дальнейшего количественного оценивания и смягчения добавленной токсичности.

7.3.5 Этические аспекты исследований токсичности

Оценка машинного перевода глубоко изучена с точки зрения оценки качества, как с использованием автоматических, так и человеческих подходов к аннотации. Недавно в области ответственного искусственного интеллекта начинается развитие ответственной оценки MT, мотивированной более общей областью ответственного искусственного интеллекта. Эта оценка направлена на измерение справедливости, этических и социальных аспектов наших технологий. Например, она позволяет нам количественно измерять токсичность или предвзятости, которые наша модель сохраняет, генерирует или потенциально усиливает; например, Blodgett et al. (2020); Costa-jussà (2019); Renduchintala и Williams (2021). Среди различных альтернатив ответственной оценки мы начали с вызова токсичности, потому что он значительно влияет на вредное содержание в интернете (см. раздел 2). В этом подразделе мы обсуждаем этические вопросы, связанные с обнаружением токсичности в переводах.

Неожиданное использование списков токсичности. Не следует предполагать, что наши списки могут использоваться для модерации контента или подавления машинно-генерируемого токсичного языка. Как и в случае с человеческими переводчиками, мы считаем, что системы машинного перевода должны оставаться верными входному тексту, поэтому мы фокусируемся исключительно на обнаружении добавленной токсичности. Даже информативные и образовательные источники, которые содержат в себе некоторую степень токсичности или другие токены, которые могут быть не токсичны во всех контекстах, но будут сопоставлены нашими списками; например, некоторые страницы Википедии содержат описания сексуальных актов, а другие содержат лексику, описывающую интимные части тела в целях предоставления информации о биологии человека и здоровье. Однако мы признаем, что наши списки токсичности могут использоваться для других целей, отличных от наших, или противоположных им. Они могут быть использованы для информирования атак на классификаторы токсичности (Kurita et al., 2019), или для реализации политик, направленных на наблюдение или подавление токсичного языка. Признавая потенциальное непреднамеренное использование наших списков, мы остаемся внимательными к вероятной возможности того, что аналогичные списки могли уже быть отдельно собраны для упомянутых целей теми сущностями, которые не публично их раскрывали.

Предвзятости в построении списка. Вероятно, что сами списки токсичности включают предвзятости из-за ошибок, упущений или недостаточно разнообразных культурных фонов (Davidson et al., 2019; Gehman et al., 2020; Ross et al., 2017; Sap et al., 2019). Мы предполагаем две основные причины этих предвзятостей: (1) неоднозначности, свойственные самому токсичному языку, и (2) культурные предубеждения, которые могут быть внесены на любом этапе процесса построения списка, поскольку наши списки в основном переведены с английского.

Во-первых, мы обсудим неоднозначности в самом токсичном языке. Из-за дискомфорта или потенциального вреда, который может вызвать токсичный язык, всегда были попытки ограничить его использование в социальных взаимодействиях. Это, в свою очередь, привело социальных актеров к созданию новых средств выражения для обхода цензуры или избегания осуждения; метафоры и намеки были бы хорошими представителями таких средств выражения. Например, общие схемы лингвистического маскирования включают использование животных имен для оскорблений или ссылок на части тела. Поэтому многие слова и фразы, которые являются частью токсичного языка, также являются обычными словами, которые могут использоваться без вредных намерений. Напротив, то, что сейчас может считаться очень специфичными техническими терминами (но которые использовались более широко в разные времена в нашей истории), известно, что иногда обретают токсичное значение (например, существительное slag в английском языке). Эти лингвистические явления вызывают неоднозначности, которые в свою очередь приводят к ошибкам перевода. Для предотвращения недоразумений и разрешения неоднозначностей последние шаги по смягчению включали создание одночасового тренировочного сеанса для переводчиков и поиск более точных n-грамм в качестве замены неоднозначных одиночных токенов. Периодическое итеративное тестирование на разнообразных входных текстах, а также качественный и количественный анализ результатов, позволяют более точно заменять n-граммы, хотя ясно, что такой метод непрерывного улучшения в конечном итоге достигает точки убывающей отдачи.

Кроме известных неоднозначностей, присущих токсичному языку, культурные предвзятости также могут проникать в списки токсичности как при создании эталонного списка, так и в процессе перевода. Культурные предвзятости могут быть вызваны тем, что первоначальные черновики практически всех списков были переводами с английского (в основном американского английского). Некоторые английские записи не имеют четких прямых эквивалентов в других языках. В таком случае могут возникать ошибки и предвзятости, когда ожидания не ясно сформулированы, и переводчики пытаются предоставить переводы для всех записей любыми способами. В дополнение к уточнению ожиданий эту проблему можно смягчить, обратившись к переводчикам на промежуточном этапе выполнения переводческой задачи с просьбой указать, считают ли они, что прямой эквивалент может быть найден в их родном языке, что помогает им помнить, что они должны воздерживаться от предоставления переводов, если это не так. С другой стороны, некоторые токсичные элементы в языках, отличных от английского, не имеют четких прямых эквивалентов в английском, что

дополнительно увеличивает разрыв в покрытии между списками. Такой разрыв оказался более сложным для сглаживания. Неограниченные добавления создают шум, а ограниченные - трение. Это ставит в перспективу эффективность альтернативы созданию списков через перевод эталонного списка, что было бы создание списков исключительно путем сбора свободных предложений.

Наконец, есть также потенциальные предвзятости в процессе перевода. Предвзятости могут вноситься неравным доступом к переводчикам и консультантам по различным языкам. Для языков с высокими ресурсами доступно множество переводчиков и консультантов, которые конкурируют на рынке языковых услуг. К сожалению, это не относится к языкам с низкими ресурсами, где доступ к меньшему числу профессионалов с разнообразным опытом и специализацией увеличивает риск предвзятостей в пользу единственного мировоззрения.

Безопасность переводчиков и консультантов. По определению, область токсичности является местом работы, где переводчики и консультанты могут чувствовать себя слишком неуютно для работы (Вельбл и др., 2021). Кроме того, независимо от готовности работать с потенциально токсичным языком, мы осознаем, что производство токсичного языка, будь то через перевод, может восприниматься неодинаково в разных частях мира. В странах, где свобода слова обычно защищена законом, токсичный язык может восприниматься как оскорбительный, вызывающий негативные эмоции или вредный, и его нежелательное использование может влечь за собой последствия в определенных местах (например, на рабочем месте), но теоретически каждый может добровольно работать над сбором, курированием, переводом или анализом токсичного языка. Мы не можем с уверенностью предполагать, что так и происходит во всех других частях мира. Имея это в виду, переводчики и консультанты осведомлены о том, что их работа содержит токсичный язык, и могут отказаться от нее.

7.4 Заключение

Оценка играет ключевую роль в разработке систем машинного перевода и позволяет потенциальным пользователям оценить, насколько качество перевода достаточно хорошее для их целей. Для разработки систем мы полагаемся на автоматические метрики. Из-за разнообразия языков, среди которых некоторые лишены видимых индикаторов границ слов, мы вынуждены выходить за рамки привычной практики использования метрик BLEU. Мы используем две метрики: spBLEU, модификацию BLEU, использующую модель sentencepiece для токенизации любого текста на любом языке, и chrF++, работающую в основном на уровне символов.

Хотя автоматические оценки являются важным инструментом, они не обеспечивают одного важного аспекта: предоставление осмысленных оценок, позволяющих принимать решения о внедрении. Поэтому мы добавляем человеческую оценку в качестве финального этапа оценки системы. Однако из-за различий между человеческими оценщиками это также не гарантирует осмысленных оценок. Мы достигаем цели получения осмысленных оценок благодаря новой и более четко определенной метрике оценки (XSTS) и использованию калибровочного набора, который позволяет не только корректировать сырые оценки от оценщиков, работающих с одной парой языков, но и получать согласованные оценки по всем языковым парам.

Кроме того, мы обеспокоены о вредоносном контенте, генерируемом нашими моделями перевода. Мы разработали новые методы для обнаружения токсичности, то есть переведенного контента, содержащего токсичные слова, которых не было в исходном тексте. Благодаря ресурсу, созданному для всех 200 языков инициативы NLLB, мы смогли расширить эту работу на все эти языки: списки токсичных терминов. Вооруженные этим ресурсом, мы создали классификаторы для обнаружения и смягчения токсичности в наших переводах.

8. Объединение всего этого

Мы видели в предыдущих разделах, как различные области междисциплинарных исследований решали разные аспекты проблемы масштабирования перевода, ориентированного на человека, для языков с ограниченными ресурсами. Укрепленные нашими основополагающими принципами, описанными в разделе 2.2, мы объединяем все междисциплинарные работы из предыдущих разделов и соединяем их таким образом, чтобы улучшить производительность мультиязычных систем машинного перевода для языков с ограниченными ресурсами из малообслуживаемых сообществ. Мы делаем все это прозрачно, делясь нашими трудностями и успехами. Чтобы сообщество могло воспользоваться и развивать инициативу NLLB, мы открыто публикуем все наши оценочные бенчмарки, модели, наборы данных, код для обучения и вывода, а также другие артефакты моделирования и данных. Руководимые нашим принципом отражения, мы надеемся, что обмен нашей работой и инструментами поможет сообществу изучить текущие практики и улучшить их там, где мы терпим неудачи, в стремлении к глобальной цели оставить без внимания ни один язык.

В следующих разделах мы описываем, как мы объединяем наши различные наборы данных и улучшения из предыдущих разделов для создания массштабной мультиязычной модели машинного перевода NLLB-200, охватывающей 202 языка. Мы сравниваем производительность NLLB-200 на Flores-200 с использованием как автоматических метрик, так и результатов человеческой оценки, и демонстрируем, что мы достигаем ведущих показателей. Мы анализируем токсичность в генерациях NLLB-200 и изучаем, насколько методы фильтрации данных и токсичности могут помочь её снизить. Мы также сравниваем производительность NLLB-200 на Flores-101 и нескольких общих бенчмарках машинного перевода, показывая, что NLLB-200 хорошо обобщается и достигает конкурентоспособной производительности. Затем мы обсуждаем дистилляцию моделей и то, как некоторые из наших дистиллированных моделей предоставляют переводы для Википедии. После этого мы рассматриваем транслитерацию и диалектные переводы, а также те области, в которых мультиязычные системы могут улучшиться. Наконец, мы обсуждаем экологическое влияние NLLB-200 и последствия принятия решений, направленных на оптимизацию вычислительных ресурсов при масштабировании.

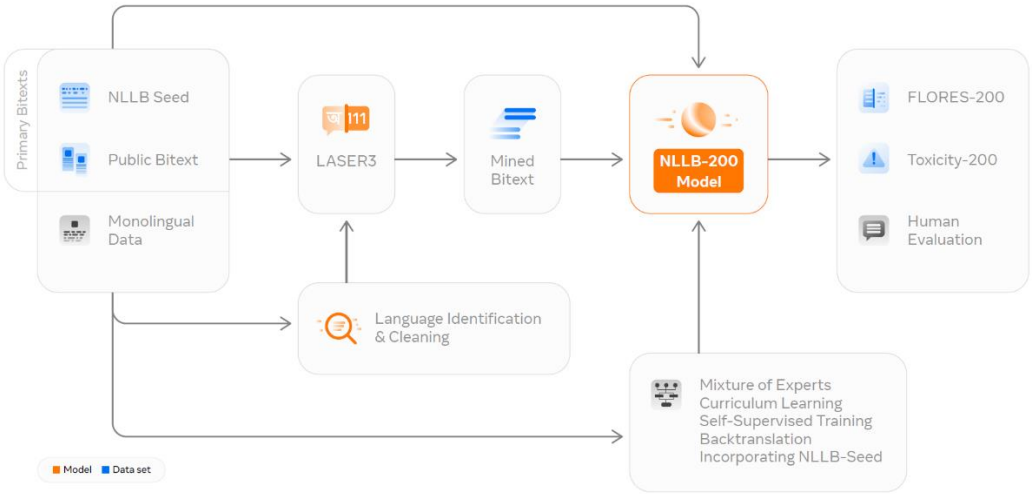


Рисунок 32: **Объединение всех компонентов**: Мы описываем, как создаем нашу конечную модель NLLB-200, используя все наборы данных и техники моделирования, которые мы создали.

8.1 Подготовка данных

Мы описываем, как мы агрегируем различные источники данных, обсужденные в предыдущих разделах, и дополняем их дополнительными данными через обратный перевод. Мы экспериментально демонстрируем преимущества использования масштабных добытых и обратно переведенных данных для значительного улучшения производительности перевода для языков с ограниченными ресурсами. Весь наш процесс кратко представлен на Рисунке 33.

8.1.1 Обучение токенизатора для 200+ языков

Для представления более чем 200 языков No Language Left Behind мы обучили новый токенизатор SentencePiece (SPM; Kudo and Richardson, 2018). Для обучения этой модели SentencePiece мы отобрали общее количество 100 миллионов предложений из основных битекстовых корпусов. Учитывая, что большинство языков в NLLB являются языками с ограниченными ресурсами (150), равномерное отборное сэмплирование приведет к перераспределению текстов высокоресурсных языков и недооценке низкоресурсных языков, что приведет к излишнему фрагментированию текстов на языках с ограниченными ресурсами. Для устранения этого мы применяем температурное сэмплирование (с температурой T = 5), которое эффективно снижает отбор высокоресурсных языков и увеличивает отбор низкоресурсных языков. Это приводит к более сбалансированному распределению сэмплов по всем языкам.

Для проверки качества SPM мы сначала изучаем уровень неизвестных токенов (<unk>) для каждого языка. Мы замечаем, что даже после использования высокой температуры для сэмплирования, определенные языки, такие как zho_Hans, zho_Hant и yue_Hant, имеют более высокие уровни ошибок <unk>, из-за очень большого набора символов их письменности. Для компенсации мы дополнительно увеличиваем отбор этих конкретных языков в 5 раз во время обучения. С этими модификациями уровень ошибок <unk> для всех языков составляет менее 1%. Еще одним важным фактором для качества является уровень токенизации, или среднее количество токенов на предложение для каждого языка (Mielke и др., 2021).

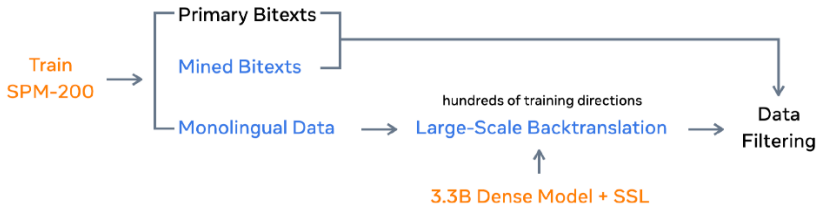


Рисунок 33: **Формирование окончательного набора данных.** Мы показываем, как мы объединяем различные источники данных, техники аугментации данных и фильтрации данных. Созданные нами наборы данных обозначены синим цветом, а модели - оранжевым.

Так как SentencePiece идентифицирует подсловные единицы на основе языковой перплексии (приблизительно, частотности), языки с недостаточным количеством данных имеют более высокий уровень токенизации по сравнению с языками с высокими ресурсами, что приводит к почти символьной модели для этих языков. Это усложняет моделирование, особенно для зависимостей на больших расстояниях и для синтеза слов из токенов, близких к уровню символов. Исходя из этих двух факторов, мы выбираем словарь размером 256,000 для нашей модели SentencePiece, чтобы обеспечить достаточную емкость для представления широкого спектра и большого количества языков, которые мы охватываем. Поскольку мы достигаем приемлемого качества токенизации с размером словаря 256k, мы не обучаем модели SentencePiece с еще большими размерами словаря (например, 384k или больше), поскольку увеличение размера словаря значительно увеличивает количество параметров модели.

Для оценки с использованием spBLEU мы используем эту модель SPM-200 в качестве токенизатора для лучшей поддержки языков Flores-200. Мы открыто распространяем эту модель SentencePiece вместе с набором данных Flores-200.

8.1.2 Наборы данных

Для обучения наших систем мы используем три различных типа данных, перечисленных ниже:

Основной битекст.** Мы используем набор общедоступных параллельных корпусов из различных источников, включая NLLB-Seed (раздел 4.2). Мы добавляем общее количество 661 наборов основных битекстовых данных. Мы выбираем все наборы, сосредоточенные на английском языке, если они доступны, а также добавляем наборы, не сосредоточенные на английском языке, если в них присутствует язык с низкими ресурсами в качестве исходного, целевого или обоих. В таблице 52 в Приложении E.1.1 предоставлена дополнительная информация о списке общедоступных битекстовых корпусов, используемых для обучения. Мы используем термин «Основной», чтобы обозначить этот тип данных.

Добытый битекст. Мы используем битекстовые корпуса, полученные путем массового добычи битекстов, подробно описанной в разделе 5.3. Мы добавляем данные о добыче для общего числа 784 направлений. Среди них все направления, сосредоточенные на английском языке, и подмножество направлений, не сосредоточенных на английском языке. Данные о добыче, не сосредоточенные на английском языке, эффективно улучшают производительность мультиязычных систем перевода (Fan et al., 2020). Однако более 200 языков подразумевает более 40,000 направлений добычи битекстов, что может быть вредным, так как некоторые пары не будут иметь высококачественных добытых битекстов. Для подселекции на основе предполагаемого качества мы первоначально выбираем направления, у которых коэффициент ошибки xsim меньше 5. Как дополнительное ограничение, мы добавляем данные о добыче в основном для пар, содержащих языки с низкими ресурсами в рамках определенной языковой семьи или географического региона. Это несовершенное приближение, направленное на обеспечение улучшенного переноса обучения между схожими языками. В Приложении E.1.3 мы делимся полным списком направлений, добытых битекстов, используемых для обучения. Мы используем термин «Добытый», чтобы обозначить этот тип данных.

Монолингвальный текст.** Подробности о наших монолингвальных данных можно найти в разделе 5.2. Мы используем монолингвальные данные для 192 языков. Эти данные используются для самообучения и для генерации данных с обратным переводом (описано в следующем разделе). Мы используем термин "Монолингвальный", чтобы обозначить этот тип данных.

8.1.3 Масштабный обратный перевод

Данные с обратным переводом предоставляют форму слабого контроля, который имеет решающее значение для улучшения качества перевода языков с низкими ресурсами. Как мы заметили в разделе 6.4.1, комбинация данных с обратным переводом, полученных из нескольких источников, улучшает производительность модели перевода за счет увеличения разнообразия обратного перевода. Следуя этому результату, мы генерируем данные с обратным переводом из двух моделей: (1) мультиязычной модели машинного перевода (MmtBT) и (2) набора двуязычных статистических моделей машинного перевода (SmtBT). Далее мы опишем, как мы выбираем языковые пары для обратного перевода.

Идентификация направлений обратного перевода.** Несмотря на свою эффективность, обратный перевод является вычислительно затратным, так как требует обучения массово мультиязычных, а также двуязычных моделей и генерации переводов для десятков миллионов предложений по каждому направлению. Мы опишем, как мы определяем, какие направления перевода принесут наибольшую пользу от увеличенного объема данных с обратным переводом.

Сначала мы обучаем базовую мультиязычную модель машинного перевода (3.3B параметров, плотная) на наборе данных, состоящем из всех основных пар битекстов и пар, сосредоточенных на английском языке, добытых данных. Мы обучаем эту модель в многозадачной конфигурации, состоящей как из задач MMT, так и задач SSL, чтобы создать как можно более сильную модель. Мы выяснили, что крайне важно выполнять обратный перевод с использованием максимально качественной модели, так как качество получаемых обратных переводов сильно коррелирует с производительностью модели, используемой для генерации данных с обратным переводом (подробности см. на рисунке 43). В разделе 6.3 мы заметили, что обучение на самообучающихся задачах в дополнение к задаче MMT в многозадачной конфигурации улучшает производительность, особенно при обучении на данных Primary+Mined (см. раздел 8.2.1). Улучшения в направлениях xx-eng_Latn значительнее по сравнению с направлениями eng_Latn-xx для языков с низкими ресурсами. Это является дополнительным преимуществом для модели при генерации обратных переводов для eng_Latn-xx.

После обучения вышеупомянутой базовой модели мы выбираем подмножество языков, для которых генерируем данные SmtBT. Мы выбираем каждый язык xx, для которого базовая модель достигает spBLEU < 10 в направлениях eng_Latn-xx или spBLEU < 15 в направлениях xx-eng_Latn на наборе данных Flores-200 dev. Эти пороговые значения выбраны, чтобы количество направлений обратного перевода оставалось управляемым с учетом вычислительных ограничений, и также основываются на предыдущих предварительных экспериментах, которые показали, что прирост от использования SmtBT был сосредоточен в направлениях, по которым базовая модель показывала плохие результаты. Это обычно верно для очень редких и недостаточно обслуживаемых языков. Для каждого из выбранных языков мы затем производим данные SmtBT как на английском, так и с английского.

Наконец, мы определяем направления, для которых необходимо получить данные MmtBT. Так как нейронные модели особенно эффективны в сценариях с высокими ресурсами, мы увеличиваем порог для охвата более широкого спектра языков. Мы выбираем каждое направление на и с английского, где производительность базовой модели на наборе данных Flores-200 dev ниже 30 spBLEU. В целом, используя эти критерии, мы выбрали 76 направлений, сосредоточенных на английском языке, для обратного перевода через конвейер SmtBT и 261 направление через MmtBT.

8.1.4 Стратегия фильтрации

Добавление данных с обратного перевода и добычи может значительно повысить производительность модели. Однако эти процессы обычно приводят к получению пар предложений, которые намного шумнее, чем данные, переведенные человеком. Чтобы воспользоваться такими данными и одновременно ограничить негативные эффекты шума, необходимо выполнить серию шагов по фильтрации битекста. Фильтрация корпусов используется в машинном переводе на протяжении многих лет (Koehn et al., 2018, 2019, 2020). Наш конвейер фильтрации выполняет несколько типов проверок, предназначенных для определения того, является ли данная пара данных настоящей парой перевода. Мы можем разделить их на несколько категорий.

Фильтрация LASER.** Фильтр LASER работает с добытыми данными и может удалять пары битекстов, чей LASER-скор ниже заданного порога (см. Раздел 5.3). Для всей нашей фильтрации добытых битекстов мы применяем порог 1.06 (Schwenk et al., 2021b).

Фильтрация по длине.** Фильтр по длине идентифицирует предложения, которые не соответствуют заданным максимальным и минимальным соотношениям длины. Он также отфильтровывает пары предложений с сильно перекошенными соотношениями длины, так как такие пары обычно вряд ли представляют собой реальные переводы.

Понятие длины не является простым при работе с массово мультиязычной моделью из-за присущих различий в распределении длины предложений для разных языков. Чтобы учесть это явление, мы измеряем длину предложений на основе кодовых точек Unicode и применяем языково-специфический корректирующий фактор. Для данного языка ` фактор вычисляется как \(\beta_\ell = \frac{N_{eng\_Latn}}{N_\ell}\), где \(N_\ell\) — это общая длина валидационного набора данных Flores-200 для языка `. Конкретно, умножение всех длин на этот фактор позволяет нам выразить единый порог длины в терминах типичной длины символов на английском языке и применять скидку или штраф по длине для языков с обычно более короткими или длинными предложениями соответственно. Для добытых и обратнопереведенных данных мы отфильтровываем предложения с соотношениями длины выше 9.0. Для данных обратного перевода мы дополнительно отфильтровываем предложения длиной менее 15 символов (соответствует примерно трем словам на английском языке), так как эмпирически выяснилось, что такие короткие предложения являются источником шума.

Фильтрация LID.** Еще один важный шаг — отбрасывание пар, предложения в которых не кажутся написанными на ожидаемых языках. Это можно выполнить автоматически с помощью идентификации языка (Раздел 5.1), с пороговыми значениями, выбранными соответствующим образом на основе надежности оценок LID для каждого данного языка.

Фильтр токсичности.** На основе техник, предложенных в разделе 7.3, мы внедряем фильтр токсичности. Он удаляет пары предложений, которые имеют дисбаланс токсичности, т.е., когда разница в количестве токсичных элементов, обнаруженных в исходном и целевом предложениях, превышает определенный порог. Альтернативный режим работы, поддерживаемый фильтром, заключается в удалении всех пар, в которых одно или оба предложения содержат токсичные элементы выше заданного порога, независимо от их относительной разницы. Дополнительные подробности приведены в разделе 7.3.3.

Дедупликация.** Ли и др. (2021) демонстрируют, что дедупликация данных для обучения критически важна для обучения больших языковых моделей. Дедупликация важна и для машинного перевода, поэтому на заключительном этапе фильтрационного конвейера мы удаляем дубликаты. Чтобы определить, являются ли два текста дубликатами, мы применяем процесс нормализации, который удаляет знаки препинания и непечатаемые символы, а затем заменяет все цифры. Фильтрация может удалить пары-дубликаты, определяемые как случаи, когда два элемента данных имеют идентичные исходный и целевой тексты; или дубликаты источника/цели, т.е. элементы данных, которые имеют одинаковую исходную сторону, но могут иметь разные цели и наоборот. Дедупликация источников полезна для данных с обратным переводом, чтобы выявить случаи, когда модель обратного перевода декодирует одну и ту же последовательность независимо от её входных данных. Дедупликация целей полезна для добытых данных, где одинаковые исходные предложения выровнены с несколькими целевыми предложениями.

Эффект фильтрации подробно обсуждается в разделе 7.3.3, где мы проводим эксперименты с обучением двуязычных моделей для шести направлений на отфильтрованных и неотфильтрованных данных, сравнивая их производительность, а также количество генерируемой галлюцинированной токсичности. Результаты в таблице 27 подтверждают, что фильтрация приводит к улучшению производительности, а также к уменьшению галлюцинированной токсичности. Наша конфигурация фильтрации для различных источников данных доступна вместе с другими нашими скриптами для обучения.

8.1.5 Влияние использования различных источников данных на производительность

Мы ожидаем кумулятивной выгоды от комбинирования различных источников данных. В этом разделе мы эмпирически исследуем эту гипотезу.

Экспериментальная установка.** Мы обучаем плотные модели Transformer с 3,3 миллиардами параметров (энкодер-декодер) с размером модели 2048, размером FFN 8192, 16 головками внимания и 48 слоями (24 энкодера, 24 декодера) для этих экспериментов по исключению данных. Мы обучаем эти модели на трех наборах данных: (1) Primary, (2) Primary+Mined и (3) Primary+Mined+MmtBT+SmtBT, чтобы сравнить кумулятивные улучшения, возникающие при добавлении каждого источника данных. Все модели обучаются в общей сложности на 300 тысяч итераций, и мы сообщаем результаты по контрольным точкам с наилучшим показателем chrF++.

Результаты.** На рисунке 34 мы демонстрируем влияние добавления различных источников данных на данные Primary. Мы агрегируем результаты по типу языковой пары и уровню ресурсов. Мы наблюдаем, что по всем языковым парам производительность значительно улучшается при добавлении добытых данных и еще больше при добавлении данных с обратным переводом MmtBT+SmtBT. Сосредоточив внимание на уровне ресурсов, мы наблюдаем, что языки с низким уровнем ресурсов улучшаются больше по сравнению с языками с высоким уровнем ресурсов. Это неудивительно, так как для языков с высоким уровнем ресурсов уже имеется значительное количество данных Primary битекста, доступных публично.

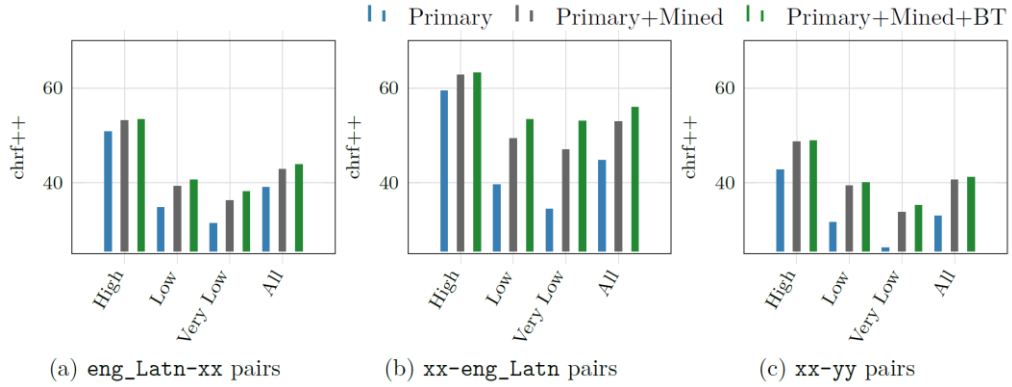


Рисунок 34: **Сравнение производительности модели при обучении на данных из различных источников**. Мы наблюдаем значительные улучшения при добавлении добытых данных и данных с обратным переводом MmtBT+SmtBT для всех типов языковых пар и уровней ресурсов.

Влияние добычи и обратного перевода на языки с очень низкими ресурсами.** При более глубоком анализе результатов мы исследуем, как источники данных, полученных путем добычи и обратного перевода, влияют на языки с очень низкими ресурсами. Мы определяем языки с очень низкими ресурсами как те, которые имеют менее 100 тысяч уникальных пар предложений среди всех языковых пар, доступных в публичных битекстовых корпусах. В целом, предлагаемые нами методы добычи и обратного перевода значительно улучшают направления для языков с низкими и очень низкими ресурсами (см. рисунок 34). Особенно заметно, что направления с очень низкими ресурсами на английский язык улучшаются на +12,5 chrF++ с добытыми данными и на +6,1 chrF++ с дополнительными данными обратного перевода, с общим улучшением на +18,6 chrF++.

Аналогично, мы наблюдаем, что направления с английского языка улучшаются на +4,7 chrF++ при добавлении добытых данных и на +1,9 chrF++ при добавлении данных обратного перевода, с общим улучшением на +6,6 chrF++. Для не англоцентричных пар мы видим улучшение на +7,5 chrF++ при добавлении добытых данных и на +1,4 chrF++ при добавлении данных обратного перевода, с общим улучшением на +8,9 chrF++. Эти результаты показывают, что наши улучшения в добыче битекста и обратном переводе увеличивают как количество, так и качество данных для языков с низкими ресурсами, которые часто остаются без внимания или исключены существующими системами перевода.

8.1.6 Набор данных для 200 языков

Мы наблюдали преимущества как добычи, так и обратного перевода для языков с низкими ресурсами. Объединив все источники данных, мы подготовили наш окончательный набор данных, охватывающий 202 языка. Набор данных включает основной битекст для 661 языковой пары, добытый битекст для 784 языковых пар и 261 направление для битекста с обратным переводом. В общей сложности, имеется 1220 языковых пар или 2440 направлений (xx-yy и yy-xx) для обучения. Эти 2440 направлений суммируются в более чем 18 миллиардов пар предложений. На рисунке 35 показано распределение образцов по 1220 языковым парам — большинство пар имеют менее 1 миллиона предложений и относятся к направлениям с низкими ресурсами.

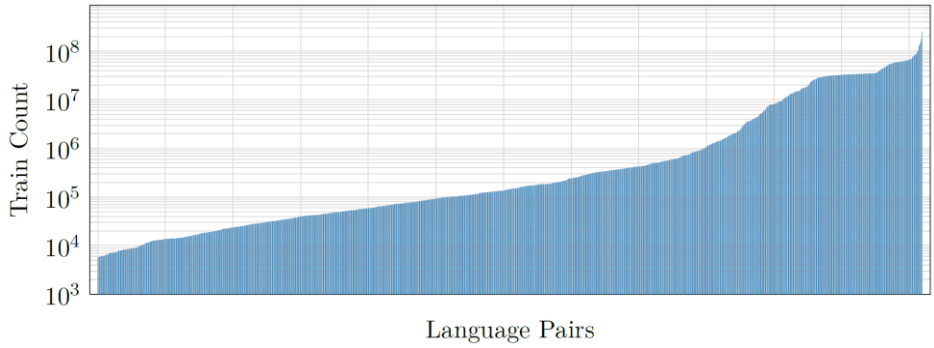


Рисунок 35: **Распределение количества тренировочных пар предложений по 1220 языковым парам в нашем наборе данных**. Мы наблюдаем, что большинство пар имеют менее 1 миллиона предложений и относятся к направлениям с низкими ресурсами.

8.2 Подготовка модели

В предыдущем разделе мы обсуждали, как улучшить количество и качество данных с помощью добычи, обратного перевода и фильтрации, что приводит к значительным улучшениям в производительности модели на языках с низкими ресурсами. В этом разделе мы обсуждаем, как мы масштабируем и адаптируем архитектуру нашей модели и процедуру обучения для создания многоязычных моделей машинного перевода для более чем 200 языков и тысяч направлений перевода. Обучение больших моделей в условиях массового многоязычия представляет собой сложную задачу из-за крайнего дисбаланса данных между языковыми парами, как показано на рисунке 35, и различного уровня сложности перевода. Цели обучения в многоязычных условиях имеют сложную и неизвестную динамику и часто конкурируют друг с другом из-за интерференции градиентов (Wang et al., 2020c). Языковые пары с низкими ресурсами быстро переобучаются, в то время как языковые пары с высокими ресурсами обычно выигрывают от более длительного обучения. В целом, эти противоречивые динамики обучения делают задачу оптимизации сложной.

Мы решили некоторые из этих задач в разделе 6.2, показав, как модели Sparsely Gated Mixture of Expert с различными стратегиями регуляризации и курсового обучения помогают улучшить производительность массово многоязычных моделей машинного перевода, особенно для языков с низкими ресурсами. В разделе 6.3 мы продемонстрировали, как можно использовать монолингвальные данные для улучшения многоязычного машинного перевода через самоконтроль в форме дополнительной задачи автокодировщика для устранения шума во время обучения. В разделе 6.4 мы рассмотрели другой способ использования монолингвальных данных через масштабный обратный перевод.

Теперь мы применяем эти стратегии на полном тренировочном наборе данных, как описано в предыдущем разделе. Сначала мы анализируем преимущества самоконтроля (SSL) с задачей автокодировщика для устранения шума (DAE) при обучении с данными обратного перевода и без них. Это помогает нам понять, помогает ли SSL дополнительно к добыче и обратному переводу, поскольку все подходы используют одни и те же монолингвальные данные. Затем мы применяем наиболее перспективные стратегии регуляризации и курсового обучения из раздела 6.2 и обучаем модели Sparsely Gated Mixture of Expert (MoE) на полном наборе данных. Мы анализируем влияние частоты слоев MoE в модели. С наилучшей частотой слоев MoE и стратегией регуляризации мы затем анализируем влияние введения языковых пар с курсовым обучением, основанным на свойствах переобучения каждой языковой пары. На основе этих экспериментов мы предлагаем архитектуру модели и рецепт обучения для создания нашей окончательной модели, NLLB-200, массово многоязычной модели машинного перевода, охватывающей 202 языка и способной переводить более 40 тысяч направлений языковых пар.

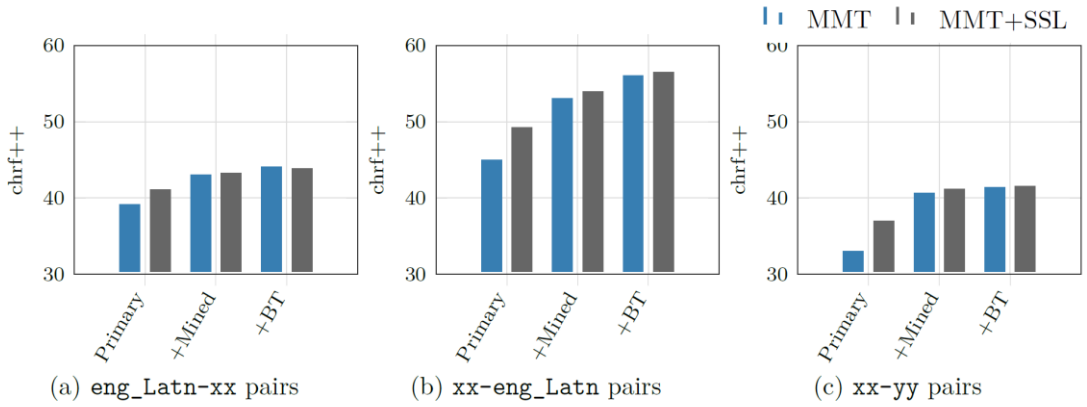


Рисунок 36: **Сравнение производительности только MMT и MMT+SSL в многозадачном режиме** в совокупности по направлениям "на английский" и "с английского", при обучении на разных источниках данных. Мы сравниваем по показателям chrF++. Мы наблюдаем, что по мере добавления все большего количества данных из добычи и/или обратного перевода, улучшения от задачи SSL начинают уменьшаться.

8.2.1 Помогает ли самообучение на основе самообучения (SSL) поверх добычи данных и обратного перевода?

Используя крупномасштабную добычу двуязычных текстов, как обсуждалось в разделе 5.3, мы уже используем подмножество монолингвальных предложений на малообеспеченных языках, которые можно успешно сопоставить с предложениями на других языках. Как подробно описано в разделе 6.3, цели самообучения могут помочь дополнительно использовать монолингвальные данные для улучшения производительности на малообеспеченных языках (Bapna et al., 2022; Liu et al., 2021a; Ma et al., 2021). В разделе 8.1.3 мы видели, что самообучение приносит значительные улучшения модели многоязычного нейронного машинного перевода, обученной на первичных двуязычных и добытых данных, и эта модель затем используется для генерации данных обратного перевода. Важно отметить, что те же самые монолингвальные данные снова используются для генерации обратного перевода. Переводы на малообеспеченные языки значительно улучшаются, когда мы добавляем добытые двуязычные тексты и двуязычные тексты обратного перевода, как видно из результатов в разделе 8.1.5. Здесь мы показываем некоторые эксперименты и связанные с ними выводы, чтобы понять, насколько больше улучшений может принести самообучение при комбинировании с обучением на добытых, а также на данных обратного перевода.

Экспериментальная установка. Мы используем настройку, аналогичную экспериментам в разделе 8.1.5 на различных наборах данных и обучаемся с самообучением (SSL) в многозадачном режиме и без него. Используемые наборы данных: (1) Primary, (2) Primary + Mined и (3) Primary + Mined + MmtBT + SmtBT. Мы следуем аналогичной настройке, как в разделе 6.3 для многозадачного обучения и обучаем плотные модели размером 3.3B в течение 300,000 обновлений. Мы сравниваем модели на лучших контрольных точках chrF++.

Результаты. На рисунке 36 мы наблюдаем, что при использовании дополнительной задачи самообучения с автоэнкодером устранения шумов (DAE) на монолингвальных данных в дополнение к задаче многоязычного машинного перевода (MMT) улучшения уменьшаются по мере добавления добытых данных и становятся незначительными, когда мы добавляем данные обратного перевода MmtBT+SmtBT. Это демонстрирует, что добыча и обратный перевод предоставляют качественные дополненные двуязычные тексты из монолингвальных данных, что улучшает переводы на малообеспеченные языки. При комбинировании с обратным переводом многозадачное обучение MMT+SSL начинает давать худшие результаты по направлениям eng_Latn-xx и не дает улучшений по направлениям xx-eng_Latn и xx-yy. Основываясь на этих результатах, для нашего финального обучения модели мы не используем SSL и просто обучаемся на задаче MMT. Тем не менее, безусловно, могут быть ситуации, когда невозможно получить достаточно качественные добытые или обратные переведенные данные, и некоторая форма самообучения на более качественных и/или относительно больших монолингвальных корпусах может быть перспективной. Лучшая стратегия использования монолингвальных данных является активной областью исследований, и в будущем работа может привести к более глубокому пониманию относительных преимуществ добычи данных, обратного перевода, самообучения и лучших способов их комбинирования.

8.2.2 Масштабирование архитектуры модели

Готовясь к обучению на финальном наборе данных, включающем 202 языка и более 18 миллиардов пар предложений, и 2440 языковых направлений, нам необходимо масштабировать наши модели, чтобы обеспечить достаточную емкость для качественного моделирования всех языков и языковых направлений. Мы умеренно увеличиваем размер плотной модели до 3.3B параметров. Помимо этого, мы масштабируем размер модели, используя стратегию условного вычисления модели со слабо соединенными экспертами (Sparsely Gated Mixture of Expert, MoE) по двум причинам: (1) Добавление экспертной емкости помогает языкам с высокими ресурсами за счет увеличения емкости модели, а также помогает малообеспеченным языкам, минимизируя интерференцию с неродственными языками. (2) Вычислительная нагрузка (FLOPs) моделей MoE не увеличивается пропорционально количеству параметров. Однако большие модели, особенно MoE, склонны к быстрому переобучению на малообеспеченных направлениях, поэтому мы применяем стратегии регуляризации и обучения по учебной программе, обсужденные в разделе 6.2, чтобы оптимизировать сложную динамику обучения при обучении 2440 различных языковых пар с различными уровнями ресурсов, различной сложностью и различной связанностью с другими языковыми парами.

Слои Sparsely Gated MoE. Теперь мы опишем, как мы интегрируем слои Sparsely Gated MoE в наши модели. По умолчанию существующая литература (Artetxe et al., 2021; Fedus et al., 2022; Lepikhin et al., 2020) заменяет слой Feed Forward Network (FFN) на каждом втором блоке Transformer на слой MoE. Количество экспертов выбирается в зависимости от того, насколько мы хотим увеличить емкость модели MoE и характера задачи, на которую обучаемся. Масштабирование числа экспертов в моделях MoE сопровождается увеличением стоимости распределенной коммуникации во время обучения из-за дорогостоящей примитивной операции All-to-All, которая выполняется дважды (диспетчеризация и объединение) для каждого слоя на каждом обновлении модели (Lepikhin et al., 2020). Еще одним недостатком является то, что примитивная операция All-to-All масштабируется сублинейно с размером распределенного мира (Lepikhin et al., 2020). Для улучшения эффективности коммуникации нашей модели в процессе обучения мы используем 128 экспертов. Это решение основано на том, что мы наблюдаем, что 32 эксперта работают так же хорошо, как 64 эксперта на бенчмарке из 53 языков, когда общий dropout настраивается отдельно для каждой конфигурации в разделе 6.2.

	eng_Latn-xx				xx-eng_Latn				xx-yy	Avg.
	all	high	low	v.low	all	high	low	v.low	all	all
f_{MoE}=2	44.6	**54.3**	41.2	38.8	56.0	**63.8**	53.2	52.1	41.4	47.3
f_{MoE}=4	**44.8**	**54.2**	**41.4**	**39.1**	**56.3**	**63.8**	**53.5**	**52.8**	**42.0**	**47.7**
f_{MoE}=6	44.6	54.0	41.2	39.0	56.1	63.7	53.4	52.7	41.7	47.5

Таблица 28: **Влияние частоты вставки слоев MoE**. Мы отчитываемся о баллах chrF++ на наборе данных Flores-200 на различных типах языковых пар. Для eng_Latn-xx и xx-eng_Latn мы включаем все 201 пару каждую. Для xx-yy мы случайным образом выбираем 200 направлений. Мы наблюдаем, что частота вставки слоев MoE равная 4 обеспечивает лучшую производительность по всем типам пар и в среднем.

Языковой бенчмарк, когда общая отсечка настроена отдельно для обоих настроек в разделе 6.2. Другой подход к повышению эффективности обучения заключается в снижении частоты слоев MoE путем размещения их на более широких интервалах в кодировщике и декодировщике модели. Мы экспериментируем с размещением слоев MoE на каждом втором, четвертом и шестом слое модели с 24 кодировщиками и 24 декодировщиками. Затем мы выбираем вариант с наилучшим соотношением эффективности и точности.

Выбор оптимальной стратегии регуляризации. Отсечка (Srivastava et al., 2014) является сильной техникой регуляризации для снижения переобучения в глубоких нейронных сетях. В разделе 6.2.2 мы расширяем и улучшаем общую отсечку и разрабатываем специфические техники регуляризации для слоев Mixture of Expert (MoE). Исходя из результатов и анализов вариантов, которые мы попробовали, мы выбираем стратегию регуляризации EOM, которая обеспечивает лучший баланс между производительностью/точностью и эффективностью. EOM требует на 23% меньше операций с плавающей запятой по сравнению с CMR. Хотя CMR немного лучше в терминах производительности, мы оптимизируем для эффективности обучения, а не для маргинального прироста производительности.

Экспериментальная установка. Мы сравниваем модели MoE-128 в трех настройках, варьируя частоту размещения слоев MoE: fMoE 2 f2; 4; 6g. Мы используем стратегию EOM отсечки для всех этих моделей и оставляем все остальные параметры одинаковыми. Все модели обучаются на 200k обновлений, и мы выбираем точку контроля с лучшим показателем chrF++.

Результаты. Как видно из Таблицы 28, вставка слоев MoE с интервалом каждые 4 блока Transformer показывает лучшую производительность, особенно улучшая ее в условиях очень ограниченных ресурсов. В терминах эффективности обучения модель MoE с fMoE=4 эффективнее на 28% по сравнению с fMoE=2.

8.2.3 Проектирование оптимизированного обучающего курса

Независимо от архитектуры модели, оптимальный учебный курс в условиях массового мультизадачного обучения помогает найти лучший локальный минимум и действует как сильный регуляризатор (Bengio et al., 2009; Lu et al., 2020). Существует несколько типов стратегий обучения по курсу (curriculum learning), основанных на том, как мы упорядочиваем задачи — по уровню их сложности или по их сходимости. Для массового мультиязычного машинного перевода высокоресурсные пары продолжают сходиться при более длительном обучении, в то время как низкоресурсные пары быстро переобучаются из-за меньших обучающих наборов данных и большей емкости модели. Мы предлагаем простую стратегию, в которой сначала вводятся направления с высокими ресурсами, а затем в процессе обучения вводятся пары, которые начинают переобучаться позднее. Раздел 6.2.3 указывает на выгоду от этой стратегии для направлений с низкими и очень низкими ресурсами. Исходя из наблюдений за моментами переобучения различных направлений в разные моменты обучения, мы экспериментируем с тем, как оптимально разделить языковые пары на фазы введения во время обучения.

Экспериментальная установка. Здесь мы сравниваем три стратегии обучения по курсу: (1) без курса, (2) курс с наивным разделением на ячейки на основе количества примеров в обучающих данных и (3) курс с разделением на несколько фаз на основе наблюдаемого переобучения при обучении без курса. Мы обучаем модели MoE с 128 экспертами с fMoE=4 для каждой стратегии и устанавливаем T = 300k, где T - общее количество обновлений для всего обучения. В наивной стратегии курса (2) мы разделяем языковые пары на (2a) те, у которых более 9 миллионов пар предложений в обучающем наборе, и (2b) те, у которых менее 9 миллионов пар предложений. Мы начинаем обучение с языковых пар, у которых более 9 миллионов предложений (261 языковых пар, включая 128 с низкими ресурсами). На 200k обновлений мы вводим языковые пары с менее чем 9 миллионами предложений (2179 языковых пар, включая 2010 с низкими ресурсами) и обучаем все пары в общей сложности до T обновлений. Этот порог был выбран эмпирически на основе наблюдаемой зависимости между количеством примеров обучения для каждой языковой пары и количеством обновлений до переобучения.

В варианте фазированного курса (3) мы сначала обучаем модель с вариантом (1) без курса, а затем разделяем языковые пары на n различных ячеек b0; b1; ...; bn-1 в зависимости от того, когда они начинают переобучаться. Затем мы начинаем обучение заново, вводя конкретную ячейку bi на обновлениях T - ki, где ki - медианный номер обновлений, после которого все направления в ячейке bi начинают переобучаться. В нашем конкретном случае мы устанавливаем k0 = 300k, k1 = 130k, k2 = 70k, k3 = 30k. Пары в b0 вводятся первыми в начале обучения модели, пары в b1 вводятся на шаге T - 130k, и так далее. Точный состав каждой ячейки в фазированном курсе объясняется в приложении E.1.3. Кратко, есть 822 пары, включая 673 пары с низкими ресурсами в b0. Есть 506 пар, включая 449 пар с низкими ресурсами в b1. Есть 455 пар, включая 444 пары с низкими ресурсами в b2. Наконец, есть 657 пар, включая 612 пар с низкими ресурсами в b3.

Результаты. Из Таблицы 29 мы видим, что использование курса, т.е. вариантов (2) или (3), лучше на этом массовом мультиязычном наборе данных по сравнению с использованием отсутствия курса, вариант (1). Это в противоречии с нашим наблюдением из раздела 6.2.3, где мы не заметили преимуществ курсового обучения в дополнение к MoE EOM. Это различие вероятно обусловлено обучением на наборе данных с четырьмя раза большим количеством языков и тысячами направлений, большинство из которых имеют низкие и очень низкие ресурсы и склонны к переобучению даже после применения сильных техник регуляризации. Направления eng_Latn-xx и xx-yy показывают улучшения более +0.6 chrF++, причем большинство улучшений происходит от направлений с низкими и очень низкими ресурсами. На направлении xx-eng_Latn мы видим улучшение +1 chrF++, и наиболее значительное улучшение на направлении с очень низкими ресурсами (+1.9 chrF++). Далее, мы замечаем, что в среднем стратегия четырехфазного курса показывает лучшие результаты среди двух вариантов курса, которые мы тестируем. По сравнению с наивной стратегией курса (2), фазированный курс (3) показывает значительные улучшения на направлениях с низким (+0.7 chrF++) и очень низкими ресурсами (+1.2 chrF++) для направлений xx-eng_Latn. Оба варианта показывают сравнимую производительность на направлениях eng_Latn-xx и направлениях, не сосредоточенных на английском языке. Мы подробнее проанализируем эффект фазированных учебных планов в разделе 8.5.2.

	eng_Latn-xx				xx-eng_Latn				xx-yy	Avg.
	all	high	low	v.low	all	high	low	v.low	all	all
(1) No CL	44.8	54.3	41.4	39.0	56.2	64.0	53.4	52.5	41.9	47.6
(2) Naive CL	**45.4**	**54.9**	**42.0**	**39.8**	57.2	**64.5**	54.6	54.4	**42.6**	48.4
(3) 4-Phase CL	45.2	54.6	41.8	39.5	**57.6**	64.2	**55.3**	**55.6**	**42.6**	**48.5**

Таблица 29: **Улучшения от различных стратегий обучения по курсу (CL).** Мы отчитываемся о баллах chrF++ на наборе данных Flores-200 dev для различных типов языковых пар. Для направлений eng_Latn-xx и xx-eng_Latn мы включаем все по 201 пару. Для направлений xx-yy мы случайным образом выбираем 200 направлений. Мы замечаем, что в среднем вариант (3) 4-фазного CL показывает лучшие результаты.

8.2.4 Модель на 200 языков: NLLB-200

Основываясь на экспериментах и анализах, подробно описанных в этом разделе, мы теперь подводим итоги лучшего рецепта для построения модели NLLB-200. Наша окончательная модель представляет собой трансформерную модель кодировщик-декодировщик, в которой мы заменяем слой Feed Forward Network (FFN) в каждом четвертом блоке трансформера на слой Sparsely Gated Mixture of Experts с 128 экспертами. Мы используем размерность модели 2048, размерность FFN 8192, 16 внимательных головок, 24 слоя кодировщика и 24 слоя декодировщика. Мы используем Pre-LayerNorm (Xiong et al., 2020), как описано в разделе 6.1.1. Мы используем общую отсечку 0.3, отсечку внимания 0.1 и EOM с peom=0:2. Модель имеет общее количество параметров 54.5B и FLOPs, сходные с плотной моделью на 3.3B.

Детали обучения. Мы обучаем модель на 300k шагов, используя 4-фазовый курс, описанный в разделе 8.2.3. Мы используем эффективный размер пакета 1M токенов на обновление. Максимальная длина последовательности во время обучения составляет 512 для кодировщика и декодировщика. Мы используем оптимизатор Adam (Kingma and Ba, 2015) с параметрами 1 = 0:9, 2 = 0:98 и = 10□6. Для повышения эффективности обучения мы используем плавающую точку FP16, реализованную в библиотеке fairseq, а также поддерживаем состояния оптимизатора Adam в FP16 (Dhariwal et al., 2020). Мы линейно увеличиваем скорость обучения от 10□7 до 0:002 за 8000 шагов разогрева, затем следуем обратно пропорциональному корню из скорости обучения. Потеря представляет собой перекрестную энтропию с сглаживанием меток 0.1 (Szegedy et al., 2015). Для балансировки использования экспертов мы используем дополнительную потерю балансировки загрузки, как описано в разделе 6.2, с весом 0.01. Фактор емкости каждого эксперта установлен на 2, то есть каждый эксперт может обрабатывать до 2 x T/E токенов, где T - количество токенов в мини-пакете, а E - количество экспертов. Во время генерации мы устанавливаем емкость равной размеру пакета, чтобы все токены могли быть направлены в одного эксперта при необходимости. Это полезно для языковых пар с низкими ресурсами, которые могут предпочитать небольшой подмножество экспертов. Наконец, модель обучается принимать токен исходного языка как префикс для исходной последовательности и токен целевого языка как первый входной токен для декодера, как объяснено в разделе 6.1.1.

	eng_Latn-xx	xx-eng_Latn	xx-yy	Avg.
87 languages				
M2M-100	-/-	-/-	-/-	13.6/-
Deepnet	-/-	-/-	-/-	18.6/-
NLLB-200	**35.5**/52.2	**42.3**/62.0	**25.2**/43.3	**25.6**/43.6
101 languages				
DeltaLM	26.6/-	33.2/-	16.4/-	16.7/-
NLLB-200	**34.0**/50.7	**41.2**/60.9	**23.7**/41.4	**24.0**/41.7

Таблица 30: **Сравнение на тестовом наборе FLORES-101**. Мы оцениваем по полным 10k направлениям в FLORES-101. Мы отчитываемся как о spBLEU, так и о chrF++, где это возможно. Все значения spBLEU вычислены с использованием токенизатора Flores-101 SPM. Оценки для DeltaLM взяты с доски почёта FLORES-101. Среднее значение M2M-100 и Deepnet только по 87 языкам, совпадающим с FLORES-101; также показываем производительность NLLB-200 на этом подмножестве языков. NLLB-200 превосходит предыдущие передовые модели существенно, даже при поддержке вдвое большего числа языков.

8.3 Результаты на Flores-200

Мы представляем результаты сравнения производительности NLLB-200 с существующими передовыми моделями на языках Flores-101. Мы глубже анализируем производительность на языках Flores-200, сосредотачиваясь на направленных на английский, не англоцентрических и нулевых переводах. Мы в основном показываем результаты метрикой chrF++, а также указываем spBLEU для удобства. Затем мы проводим человеческую оценку переводов NLLB-200 на Flores-200. Наконец, мы обсуждаем распространённость галлюцинативной токсичности в генерациях NLLB-200.

8.3.1 Производительность на Flores-101 и сравнение с передовыми моделями

Мы оцениваем NLLB-200 на Flores-101 для сравнения его производительности с передовыми моделями. Мы используем тот же метод оценки, что и для Flores-101, с использованием токенизатора SentencePiece (SPM) для 101 языка, предоставленного официально. В таблице 30 мы наблюдаем, что NLLB-200 превосходит ближайший передовой уровень на почти +7.3 spBLEU в среднем — улучшение на 44%. Затем мы сравниваем с другими передовыми моделями, такими как Deepnet (Wang et al., 2022) и M2M-100 (Fan et al., 2020), которые сообщают оценки для подмножества из 87 языков Flores-101. На этом меньшем подмножестве NLLB-200 снова превосходит на +7.0 spBLEU в среднем. В целом результаты показывают, что NLLB-200 значительно улучшает существующие системы несмотря на покрытие 200+ языков — в два раза больше языков (или более 30 тыс. дополнительных направлений), чем любая предыдущая работа.

Производительность на африканских и индийских языках Далее мы сравниваем производительность NLLB-200 для двух конкретных групп языков — африканских и индийских. Несколько последних работ в области обработки естественного языка сосредоточены на африканских языках (Abbott and Martinus, 2019; Adelani et al., 2022; Azunre et al., 2021c; Dabre and Sukhoo, 2022; Emezue and Dossou, 2021, 2020; Hacheme, 2021; Nekoto et al., 2020; Siminyu et al., 2021). Здесь мы сравниваем с двумя недавними работами: MMTAfrica (Emezue and Dossou, 2021) и Mafand-MT (Adelani et al., 2022). Mafand-MT использует модель M2M-100, донастроенную на наборе данных MAFAND. Для MMTAfrica мы берём максимальный балл их методов BT и BT&REC. В таблице 31 мы наблюдаем, что NLLB-200 значительно превосходит обе модели по большинству направлений eng_Latn-xx и всех направлений xx-eng_Latn. На направлениях eng_Latn-yor_Latn и eng_Latn-swh_Latn MMTAfrica немного лучше NLLB-200. NLLB-200 значительно улучшает результаты по всем остальным направлениям, благодаря мультиязычному переносу, включая обработку 55 африканских языков, а также улучшениям в данных и моделировании. В сравнении с другими работами, которые обучаются только на 6-10 африканских языках, NLLB-200 демонстрирует значительные достижения. В таблице 53 мы далее сравниваем производительность NLLB-200 на направлениях не англоцентрических африканских языков.

	eng_Latn-xx			xx-eng_Latn		
	MMTAfrica	M2M-100*	NLLB-200	MMTAfrica	M2M-100*	NLLB-200
hau_Latn	-/-	4.0/-	**33.6/53.5**	-/-	16.3/-	**38.5//57.3**
ibo_Latn	21.4/37.2	19.9/-	**25.8/41.4**	15.4/38.9	12.0/-	**35.5/54.4**
lug_Latn	-/-	7.6/-	**16.8/39.8**	-/-	7.7/-	**27.4/46.7**
luo_Latn	-/-	13.7/-	**18.0/38.5**	-/-	11.8/-	**24.5/43.7**
swh_Latn	**40.1**/53.1	27.1/-	37.9/**58.6**	28.4/56.1	25.8/-	**48.1/66.1**
wol_Latn	-/-	8.2/-	**11.5/29.7**	-/-	7.5/-	**22.4/41.2**
xho_Latn	27.1/44.9	-/-	**29.5/48.6**	21.7/48.6	-/-	**41.9/59.9**
yor_Latn	12.0/**28.3**	13.4/-	**13.8**/25.5	9.0/30.6	9.3/-	**26.6/46.3**
zul_Latn	-/-	19.2/-	**36.3/53.3**	-/-	19.2/-	**43.4/61.5**

Таблица 31: **Сравнение на тестовом наборе FLORES-101 по африканским языкам**. Мы сравниваем с двумя недавними работами MMTAfrica и M2M-100*, донастроенными на наборе данных MAFAND-MT. Мы отчитываемся о spBLEU/chrF++ и выделяем жирным лучший результат. NLLB-200 превосходит предыдущие передовые модели существенным отрывом на большинстве направлений перевода.

В последние годы индийские языки продемонстрировали значительный прогресс в мультиязычной обработке естественного языка с низкими ресурсами (Bhattacharjee et al., 2022; Dabre et al., 2021; Kumar et al., 2022; Ramesh et al., 2022). Мы сравниваем производительность перевода NLLB-200 с (a) IndicBART (Dabre et al., 2021), (b) IndicTrans (Ramesh et al., 2022), а также

коммерческими системами перевода, такими как (c) Google Translate, (d) Microsoft Translate. NLLB-200 значительно превосходит все модели на большинстве направлений. Обучающий набор данных NLLB-200 включает 25 индийских языков37, что почти в два раза больше, чем в случае (a) и (b). Улучшение производительности можно объяснить большим мультиязычным переносом, а также улучшенным качеством данных, полученных путём майнинга и обратного перевода, для семейства индийских языков.

8.3.2 Производительность на Flores-200

Мы представляем производительность NLLB-200 на полном тестовом наборе Flores-200. Для 202 языков в NLLB-200 возможно 40,602 направлений перевода в Flores-200. Сначала мы категоризируем наши результаты на высокоресурсные, низкоресурсные и очень низкоресурсные направления, а затем подробнее разделяем каждое из них на направления из английского (eng_Latn-xx), в английский (xx-eng_Latn) и не англоцентричные направления (xx-yy). Настройка оценки. Мы оцениваем NLLB-200 на 201 направлении eng_Latn-xx, 201 направлении xx-eng_Latn и 40,200 направлениях xx-yy — всего 40,602 направления. Из них 2,862 пары являются высокоресурсными, а 37,740 пар — низкоресурсными38, из которых 26,796 также являются очень низкоресурсными39. Из них наш набор данных для обучения содержит параллельные данные для 2,440 направлений, что означает, что 38,162 направления никогда не виделись моделью во время обучения (нулевое обучение). Мы представляем результаты в Таблице 33 с использованием метрик chrF++ и spBLEU.

	eng_Latn-xx					xx-eng_Latn				
	(a)	(b)	(c)	(d)	NLLB-200	(a)	(b)	(c)	(d)	NLLB-200
asm_Beng	6.9/-	-/-	-/-	24.9/-	**27.6**/35.9	23.3/-	-/-	-/-	13.6/-	**36.1**/57.8
ben_Beng	20.3/-	17.3/-	34.6/-	31.2/-	**37.9**/50.0	32.2/-	30.7/-	28.1/-	22.9/-	**41.2**/62.2
guj_Gujr	22.6/-	22.6/-	40.2/-	35.4/-	**41.4**/53.3	34.3/-	33.6/-	25.6/-	27.7/-	**46.8**/66.6
hin_Deva	34.5/-	31.3/-	**44.2**/-	36.9/-	42.6/57.3	37.9/-	36.0/-	38.7/-	31.8/-	**46.5**/66.5
kan_Knda	18.9/-	16.7/-	32.2/-	30.5/-	**40.8**/53.4	28.8/-	27.4/-	32.6/-	22.0/-	**40.1**/61.0
mal_Mlym	16.3/-	14.2/-	34.6/-	34.1/-	**38.5**/51.6	31.7/-	30.4/-	27.4/-	21.1/-	**41.9**/62.9
mar_Deva	16.1/-	14.7/-	**36.1**/-	32.7/-	31.7/48.0	30.8/-	30.0/-	19.8/-	18.3/-	**42.8**/63.8
ory_Orya	13.9/-	10.1/-	31.7/-	31.0/-	**34.6**/45.7	30.1/-	28.6/-	24.4/-	20.9/-	**44.3**/64.4
pan_Guru	26.9/-	21.9/-	39.0/-	35.1/-	**39.8**/49.0	35.8/-	34.2/-	27.0/-	28.5/-	**47.1**/66.3
tam_Taml	16.3/-	14.9/-	31.9/-	29.8/-	**36.8**/53.7	28.6/-	27.7/-	28.0/-	20.0/-	**39.5**/60.8
tel_Telu	22.0/-	20.4/-	38.8/-	37.3/-	**43.2**/55.9	33.5/-	32.7/-	30.6/-	30.5/-	**46.1**/65.5

Таблица 32: **Сравнение на тестовом наборе FLORES-101 по индийским языкам**. Мы отчитываемся о spBLEU для всех направлений и chrF++, где это возможно, и выделяем жирным лучший результат. (a) IndicBART (Dabre et al., 2021), (b) IndicTrans (Ramesh et al., 2022), (c) Google Translate, (d) Microsoft Translate. Числа для (c) и (d) взяты из работы (Ramesh et al., 2022). NLLB-200 превосходит другие системы перевода по большинству направлений. Улучшения наиболее значительны в направлениях xx-eng_Latn.

Производительность на англоцентричных направлениях. Сначала мы обсуждаем производительность на англоцентричных направлениях. Мы наблюдаем, что в среднем направления xx-eng_Latn проявляются гораздо лучше, чем направления eng_Latn-xx — это общий тренд, который мы наблюдали ранее в разделе 6. Мы предполагаем, что это связано с несколькими причинами: (1) направления xx-eng_Latn требуют от декодера декодировать на тот же язык, и английский язык является основным в обучающем наборе данных; (2) благодаря обильным и высококачественным монолингвальным данным на английском языке, майнинг битекстов и обратный перевод обеспечивают создание более качественных данных с английской стороны; и (3) неанглийская сторона Flores-200 обычно переведена человеком и может поэтому не обладать такой же свободной грамотностью, как естественный текст из-за эффекта переводизма (Zhang и Toral, 2019). Из Таблицы 33 мы видим, что высокоресурсные направления eng_Latn-xx в среднем на +13.0 chrF++ лучше, чем низкоресурсные. Для xx-eng_Latn разрыв меньше, с высокоресурсными направлениями в среднем на +9.0 chrF++ лучше. Это также демонстрирует, что направления в английский язык проявляются лучше для языков с низкими ресурсами. Одним из преимуществ модели с лучшей производительностью на направлениях в английский язык является возможность улучшения производительности на направлениях из английского с использованием обратного перевода (раздел 8.1.3). Мы дополнительно разбиваем производительность на англоцентричных направлениях на языки Flores-101 и новые языки, добавленные в Flores-200. На направлениях Flores-101 NLLB-200 достигает 50.7 chrF++ на направлениях eng_Latn-xx и 60.9 chrF++ на направлениях xx-eng_Latn. На подмножестве новых языков (Flores-200 - Flores-101) NLLB-200 достигает 39.9 chrF++ на направлениях eng_Latn-xx и 52.6 chrF++ на направлениях xx-eng_Latn. Новые языки в Flores-200 все являются низкоресурсными и в среднем имеют значительно худшую производительность, чем языки в Flores-101. В целом системы перевода все еще имеют много места для улучшений.

Производительность в направлениях, не сосредоточенных на английском и нулевой обучение. Мы затем фокусируемся на производительности NLLB-200 в направлениях, не сосредоточенных на английском языке. В оценочном наборе данных Flores-200 есть полная поддержка многие-ко-многим, что означает, что большинство оцениваемых направлений не сосредоточены на английском языке. По сравнению с моделями, оцененными на Flores-101, Flores-200 поддерживает оценку на 30 тыс. направлений, не сосредоточенных на английском языке. Мы оцениваем 40,200 направлений xx-yy, где NLLB-200 в среднем достигает 35.6 chrF++. Это значительно ниже производительности по сравнению с xx-eng_Latn (-21.2 chrF++) или eng_Latn-xx (-9.7 chrF++), потому что большинство направлений не являются надзорными, и большинство пар имеют языки, которые не лексически или синтаксически схожи, что делает перевод более сложным.

Как упоминалось ранее, NLLB-200 никогда не видел 38,162 направлений в Flores-200. На этих направлениях zero-shot модель достигает 35.4 chrF++. В сравнении с примерно 2 тыс. наблюдаемыми парами, где модель достигает 39.5 chrF++, мы видим только снижение в среднем на -4.1 chrF++ по 38 тыс. направлений zero-shot. Нулевое обучение в направлениях с низким и очень низким ресурсами в среднем хуже примерно на -5.0 chrF++ по сравнению с надзорными парами. Модель показывает разумную производительность в нулевом обучении вероятно благодаря своему массово-многоразовому характеру.

В согласии с Fan et al. (2020), мы добавили наднадзорные многоресурсные данные для обучения нашей модели, что также способствует улучшению нулевого обучения. Кроме того, мы тщательно выбирали направления многоресурсного добытого битекста, которые имеют лучшее выровненное содержание и являются схожими языками друг с другом.

Сравнение с Google Translate. Мы сравниваем производительность NLLB-200 с коммерческими системами перевода, такими как Google Translate. Поскольку эти системы являются собственностью компаний, честные сравнения сложны, поскольку существует мало информации о архитектуре модели, настройках обучения или количестве моделей. Поэтому мы предоставляем эти результаты только для приблизительного сравнения. Мы фокусируем наше внимание на направлениях с низким и очень низким ресурсами, которые пересекаются между Flores-200 и API Google Translate40 — отметим, что NLLB-200 охватывает гораздо больше языков.

В таблице 34 показано, что NLLB-200 значительно превосходит другие модели на направлениях xx-eng_Latn, на +3.3 chrF++ как для направлений с низким ресурсом, так и для очень низким ресурсом. На направлениях eng_Latn-xx NLLB-200 отстает на -2.1 chrF++ для направлений с низким ресурсом и на -1.5 chrF++ для направлений с очень низким ресурсом. В общем, несмотря на охват 202 языков, NLLB-200 показывает лучшие результаты на +0.6 chrF++ для направлений с низким ресурсом и +0.9 chrF++ для направлений с очень низким ресурсом. Дополнительные результаты на направлениях с высоким ресурсом и дополнительный анализ представлены в приложении E.2.2.

8.3.3 Оценка качества человеком

Затем мы оцениваем качество переводов с участием профессиональных переводчиков по протоколу оценки XSTS, описанному в разделе 7.

Экспериментальные настройки. Мы сравниваем две модели: NLLB-200 и базовую плотную модель с параметрами 3.3B. Обе модели обучены на одних и тех же данных и охватывают 202 языка. Чтобы понять производительность модели на разнообразных языковых парах, мы оцениваем 34 различных направления перевода. 15 направлений из английского, 15 из не-английского и 4 направления в английский язык, которые представляют многочисленные повседневные потребности в переводе, описанные в наших исследованиях с носителями языков с низким ресурсом (см. раздел 2). Каждая оценка использует 1,012 предложений из Flores-200 devtest.

Результаты. Результаты показаны на рисунке 37. В целом, NLLB-200 достигает среднего балла XSTS 4.15 на направлениях из английского и 3.75 на направлениях в английский. По сравнению с базовой плотной моделью производительность NLLB-200 сильнее. На некоторых направлениях наблюдается значительная разница, например, на направлении rus_Cyrl-tgk_Cyrl и eng_Latn-gla_Latn. Мы также замечаем, что NLLB-200 показывает лучшие результаты по сравнению с базовой моделью на всех протестированных направлениях, за исключением eng_Latn-ell_Grek, где производительность немного хуже.

8.3.4 Распространенность токсичности

Наша цель - создание качественных безопасных переводов для каждого из наших 200 языков. Изменение значения от исходного предложения не является желательным поведением, но добавление токсичности - худшее. Мы используем детекторы токсичности, предложенные в разделе 7.3 (в частности, детектор 1+ токсичного элемента), чтобы оценить добавленную токсичность в переводах NLLB-200. Мы проводим нашу оценку на наборе данных Flores-200 devtest. Обратите внимание, что в этом наборе данных крайне низкое количество реальной токсичности (только 3 токсичных элемента в наборе devtest) из-за его домена Wikimedia.

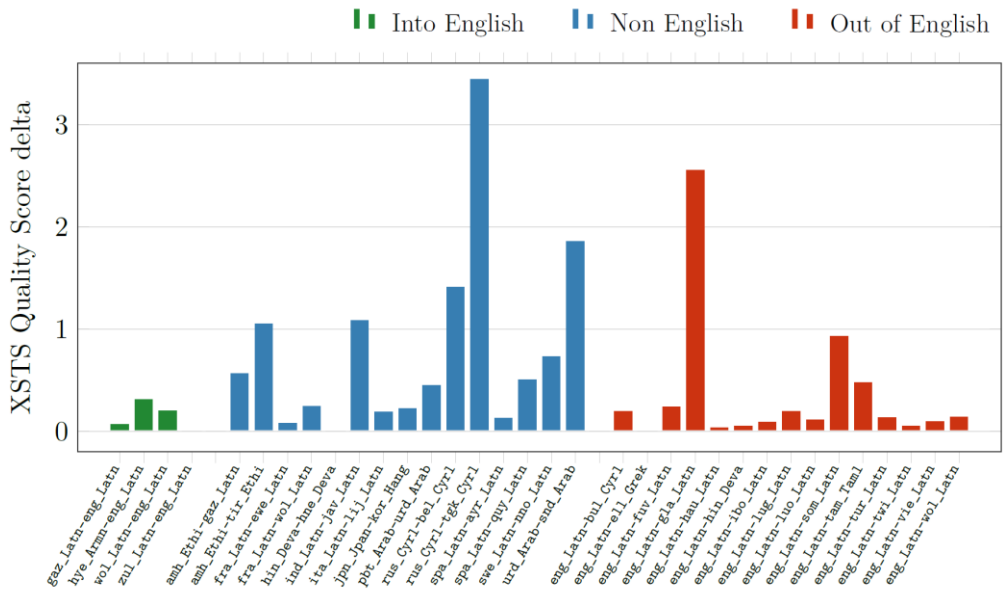


Рисунок 37: **Дельта качества оценки человеком по XSTS**. Разница между моделью NLLB-200 и плотной базовой моделью. Мы показываем средние улучшения в 0.53 по всем протестированным направлениям. Среднее улучшение для направлений в английский - 0.15, из английского - 0.36, и в среднем по направлениям не на английском - 0.81.

Используя NLLB-200, мы оценивали результаты перевода на английский и с английского. Потенциально добавленная токсичность с помощью детектора 1+ токсичного элемента была обнаружена в 130 из 201 направления eng_Latn-xx (1,636 предложений) и в 200 из 201 направления xx-eng_Latn (555 предложений). Более детальное разбиение представлено на рисунке 44 в Приложении E.3. В целом, наш анализ показывает низкое распространение потенциальной добавленной токсичности в переводах Flores-200. Однако это не означает, что наши модели в целом генерируют низкую токсичность. В других областях (например, различные темы, регистры или типы дискурса) они могут генерировать более высокие уровни токсичности. Эти области могут содержать большее количество потенциальных триггеров, таких как нестандартные токены, редкие слова или формы слов, используемые с необычными значениями

части речи (например, doff grandpa) или в необычных конструкциях (например, potbellied veterans). Кроме того, мы замечаем, что существует значительная разница в потенциальной добавленной токсичности при переводе с английского на другие языки по сравнению с переводом на английский. В переводах с английского мы имеем более чем в 3 раза больше потенциальных примеров добавленной токсичности (в среднем 0.80%), чем при переводе на английский (в среднем 0.27%). Вопрос о том, является ли эта разница результатом переизбыточного обнаружения нашими списками токсичности или действительно добавленной токсичности моделями при переводе определенных направлений, остается открытым для будущих исследований.

8.4 Обобщение за пределами домена: Производительность на доменах, не входящих в Flores-200

Далее мы оцениваем способность нашей модели к обобщению на нескольких бенчмарках машинного перевода, не относящихся к Flores-200, и сравниваем её с существующими работами. Мы также анализируем, как NLLB-200 может быть адаптирован для специализации на различных доменах и обсуждаем значительное улучшение производительности, достигаемое за счет финетюнинга внутри домена.

Наша цель в этом разделе - проверить, разрабатываем ли мы надежную универсальную систему машинного перевода, способную переводить в различных доменах. Сначала мы оцениваем способность NLLB-200 к обобщению на широкий выбор бенчмарков машинного перевода в различных доменах (новости, здравоохранение, государственные структуры и т. д.). Затем мы используем наш собственный набор данных NLLB-MD (см. Раздел 4.3) для проверки возможности передачи NLLB-200 на другие домены путем финетюнинга на небольших объемах высококачественных битекстов.

8.4.1 Публичные бенчмарки

Мы используем общедоступные и широко используемые бенчмарки оценки машинного перевода для оценки производительности NLLB-200 в доменах, отличных от текста Wikimedia в Flores-200, а также для сравнения производительности NLLB-200 с предыдущими моделями наивысшего класса. Мы выбираем общее количество 238 направлений из 8 бенчмарков. Ниже мы описываем каждый из этих бенчмарков и набор выбранных направлений.

Flores(v1): включает в себя 8 направлений, исходный набор данных Flores (Guzmán et al., 2019) пары четырех языков с низкими ресурсами (khm_Khmr, npi_Deva, pbt_Arab, sin_Sinh) с eng_Latn в домене Wikimedia.

WAT: мы выбираем 3 языка (hin_Deva, khm_Khmr и mya_Mymr), сопоставленных с английским (8 направлений) из соревнования WAT.

WMT: мы оцениваем 15 языков WMT, выбранных в работе Siddhant et al. (2020). Этот набор пересекается с 10 языками, выбранными в работе Wang et al. (2020a), и оба часто используются для тестирования моделей MMT (Kim et al., 2021; Kudugunta et al., 2021). 15 языков, сопоставленных с английским в этом наборе: (ces_Latn, deu_Latn, est_Latn, fin_Latn, fra_Latn, guj_Gujr, hin_Deva, kaz_Cyrl, lvs_Latn, lit_Latn, ron_Latn, rus_Cyrl, spa_Latn, tur_Latn и zho_Hans).

IWSLT: мы выбираем 24 направления из соревнования по переводу IWSLT. С битекстами, основанными на выровненных TED-докладах, выбранные направления взяты из различных кампаний (см. Таблицу 55 в приложении для более подробной информации по каждому направлению).

TICO: отобран из различных общедоступных источников, содержащих контент, связанный с COVID-19 (Anastasopoulos et al., 2020), этот набор данных представлен в различных доменах (медицинский, новостной, разговорный и т. д.) и охватывает 36 языков. Мы сопоставляем 28 языков с английским для общего числа 56 направлений.

Mafand-MT: африканский новостной корпус, включающий 16 языков (Adelani et al., 2022). 8 языков сопоставлены с английским (hau_Latn, ibo_Latn, lug_Latn, luo_Latn, swh_Latn, tsn_Latn, yor_Latn, zul_Latn), а 5 других языков - с французским (bam_Latn, ewe_Latn, fon_Latn, mos_Latn и wol_Latn) для общего числа 26 направлений.

	eng-xx		xx-eng	
	Published	NLLB-200	Published	NLLB-200
khm	[b]**5.9**/-	0.4/27.4	[b]10.7/-	**16.8**/36.5
npi	[c]7.4/-	**10.4**/39.0	[c]14.5/-	**29.3**/54.8
pbt	[b]9.3/-	**10.5**/34.3	[b]15.7/-	**22.0**/46.8
sin	[c]3.3/-	**11.6**/40.9	[c]13.7/-	**23.7**/49.8

(a) Flores(v1)

	eng-xx		xx-eng	
	Published	NLLB-200	Published	NLLB-200
hin	[l]22.1/-	**27.2**/51.5	[l]32.9/-	**37.4**/61.9
khm	[l]43.9/-	**45.8**/42.3	[l]27.5/-	**39.1**/61.1
mya	[c]**39.2**/-	23.5/31.5	[c]**34.9**/-	32.7/57.9

(b) WAT

Таблица 35: **Сравнение с состоянием искусства на датасетах FLORES (v1) (devtest) и тестовых наборах WAT**. Мы сообщаем оценки BLEU/chrF++, где они доступны, и выделяем жирным лучший результат. Низкоресурсные языки подчеркнуты. В каждом направлении мы отображаем лучшую модель из опубликованных работ: (b) Tang et al. (2020), (c) Liu et al. (2020) и (l) Nakazawa et al. (2019).

Autshumato: набор данных для оценки машинного перевода южноафриканских языков (McKellar, 2017) состоит из 500 предложений из данных южноафриканского правительства, переведенных отдельно четырьмя разными профессиональными человеческими переводчиками для каждого из 11 официальных языков Южной Африки. Девять из этих языков охвачены NLLB-200: afr_Latn, eng_Latn, nso_Latn, sot_Latn, ssw_Latn, tsn_Latn, tso_Latn, xho_Latn и zul_Latn. Нет стандартного разделения на

валидационную/тестовую выборки, поэтому мы используем первую половину (250 предложений, образующих 1000 пар) для валидации и вторую половину для тестирования, следуя методологии Fan et al. (2020).

MADAR: создан путем перевода выбранных предложений из базового корпуса выражений для путешествий (BTEC) (Bouamor et al., 2018). Этот корпус охватывает диалекты из 25 арабоговорящих городов, а также английский, французский и современный стандартный арабский (MDA). Мы сопоставляем 16 из этих диалектов с 8 арабскими диалектами в NLLB-200 (aeb_Arab, acm_Arab, acq_Arab, ajp_Arab, apc_Arab, ars_Arab, ary_Arab и arz_Arab) и каждый из них современным стандартным арабским (arb_Arab). Для оценки мы используем тестовую выборку MADAR shared task (corpus_6_test_corpus_26_test) (Bouamor et al., 2019).

Оценка на других наборах данных. Мы оцениваем точность переводов с помощью метрик BLEU, spBLEU и chrF++ (выбирая методологию оценки отдельных наборов данных). Для измерения BLEU мы сначала детокенизируем гипотезы. Затем для каждого корпуса и каждого языка мы приводим гипотезы и эталоны к токенизации или нормализации, используемой в текущем состоянии искусства. После токенизации гипотез и эталонов мы вычисляем BLEU. В большинстве языков мы передаем детокенизированный вывод в SacreBLEU (Post, 2018) и используем стандартный токенизатор 13a Moses. См. Таблицу 56 в приложении для подробного разбиения этапов предобработки для каждого нерегулярного направления оценки.

8.4.2 Результаты

FloresV1. На тестовых наборах данных в пределах домена Flores(v1) в Таблице 35a мы превзошли современные достижения (Liu et al., 2020; Tang et al., 2020) на всех 8 направлениях, за исключением одного (khm_Khmr-eng_Latn).

WAT. Затем, в Таблице 35b, мы оцениваем NLLB-200 на 6 направлениях из WAT. Наборы данных mya_Mymr и khm_Khmr являются частью корпуса ALT (новостные статьи), а тестовый набор hin_Deva также поступает из области новостей (IITB: Newswire). NLLB-200 превосходит современные достижения при переводе на английский, средним показателем в 3.5 балла BLEU. При переводе с английского, средний прирост составляет 3.5 балла BLEU, за исключением снижения на направлении eng_Latn-mya_Latn, что приводит средний показатель к -3 баллам BLEU.

WMT. Аналогично WAT в пределах домена, тестовые наборы WMT охватывают новостные статьи. На этом конкретном бенчмарке (см. таблицу 36a), Kim et al. (2021) достигли точности, превосходящей предыдущие достижения на порядок. По сравнению с Kim et al. (2021), NLLB-200 улучшает результаты на направлениях eng_Latn-xx примерно на 1.9 балла BLEU. Однако при переводе на английский NLLB-200 в среднем набирает на 0.6 балла BLEU меньше.

IWSLT. На 12 направлениях с высоким ресурсом языка, парных с английским в IWSLT (таблица 36b), NLLB-200 превосходит современные достижения на 21 из 24 направлений. При переводе с английского улучшение в среднем составляет 3.5 балла BLEU, а при переводе на английский - 6.3 балла BLEU.

TICO. В таблице 37 показаны результаты NLLB-200 и лучшей базовой линии, обученной в исходной статье TICO (Anastasopoulos et al., 2020). Мы видим значительный прирост точности как на низкоресурсных, так и на высокоресурсных языках. Дополнительные результаты по TICO можно найти в таблице 58.

MAFAND. Мы сравниваем в таблице 38 производительность NLLB-200 на тестовом наборе MAFAND-MT с лучшей моделью M2M-100, донастроенной в новостной домен с данными MAFAND, как сообщается в Adelani et al. (2022). NLLB-200 превосходит предыдущее современное состояние искусства на 14 из 26 протестированных направлений, доказывая, что модель хорошо обобщается на другие домены. При переводе на английский наблюдается среднее улучшение +4.0 балла BLEU.

	eng-xx		xx-eng	
	Published	NLLB-200	Published	NLLB-200
arb	15.2/-/	**34.1**/59.4	28.6/-/	**49.6**/70.3
fra	37.6/-/	**44.9**/64.4	39.4/-/	**47.3**/65.4
gaz	0.6/-/	**10.7**/44.0	2.1/-/	**35.9**/57.2
hin	6.4/-/	**46.2**/65.8	18.9/-/	**58.0**/76.2
ind	41.3/-/	**55.1**/74.8	34.9/-/	**54.3**/73.5
lin	7.8/-/	**24.6**/51.5	6.7/-/	**33.7**/54.1
lug	3.0/-/	**22.1**/48.6	5.6/-/	**39.0**/58.2
mar	0.2/-/	**16.1**/46.3	1.2/-/	**44.3**/66.9
pes	8.5/-/	**30.0**/55.6	15.1/-/	**45.5**/67.5
por	47.3/-/	**52.9**/72.9	48.6/-/	**58.7**/76.5
rus	28.9/-/	**35.7**/59.1	28.5/-/	**41.2**/65.1
spa	48.7/-/	**57.2**/74.9	46.8/-/	**57.5**/75.9
swh	22.6/-/	**34.1**/59.1	0.0/-/	**49.6**/68.1
urd	2.8/-/	**27.4**/53.3	0.0/-/	**44.7**/66.9
zho	33.7/-/	**42.0**/33.3	28.9/-/	**37.6**/61.9
zsm	6.3/-/	**52.4**/73.4	0.0/-/	**58.8**/76.1
zul	11.7/-/	**22.4**/55.1	25.5/-/	**50.6**/68.4

Таблица 37: **Сравнение по TICO**. Мы приводим оценки BLEU/spBLEU/chrF++, где они доступны, по данным Anastasopoulos et al. (2020) и выделяем жирным лучший результат. Низкоресурсные языки выделены подчеркиванием.

Однако перевод на французский (язык с высокими ресурсами) в среднем хуже на -0,9 BLEU, с значительным снижением производительности в направлении bam_Latn-fra_Latn. Перевод с английского в среднем хуже на -0,6 BLEU, а с французского - на -4,4 BLEU.

Оценка нецентрированного на английском языке на Autshumato и MADAR. Кроме того, мы оцениваем NLLB-200 на наборах данных с нецентрированными парами; MADAR для арабских диалектов и Autshumato для африканских языков, в Таблице 57 и Таблице 58 приложения.

8.4.3 Эффективная адаптация домена с NLLB-MD

Далее мы изучаем, может ли NLLB-200 эффективно переходить в другие домены и подходит ли он для общей стратегии настройки с одной задачей с небольшими объемами качественных переводов внутри домена (Adelani et al., 2022; Lee et al., 2022; Liu et al., 2021b; Tang et al., 2020).

Экспериментальная настройка. Мы проводим эксперименты с набором данных NLLB-MD (см. раздел 4.3). Он предоставляет качественные переводы в четырех областях (новости, официальные сценаризированные выступления (скриптованные), неофициальные несценаризированные выступления (чат) и здравоохранение), переводя с английского на 6 языков (из которых 5 являются низкоресурсными). Мы оставляем по 500 предложений на каждом языке для тестирования, донастраиваем на 2000 предложений и используем остальные для валидации. В каждом направлении перевода (в английский и из английского) мы донастраиваем NLLB-200 на этой единственной задаче на 50 обновлений (15-20 эпох) с коэффициентом обучения 5e-5, следуя расписанию обратного квадратного корня после разогрева в течение 10 обновлений. Мы рассматриваем два варианта донастройки NLLB-200 для новой задачи: (1) донастройка с использованием исходной целевой функции обучения (сглаженная кросс-энтропия с дополнительным регуляризационным членом для балансировки нагрузки) (см. раздел 6.2) и (2) донастройка без регуляризации, тем самым оставляя распределение нагрузки MoE неограниченным.

	eng-xx		xx-eng	
	Adelani et al. (2022)	NLLB-200	Adelani et al. (2022)	NLLB-200
hau_Latn	**15.9**/-/**42.1**	8.2/34.8	**18.2**/-/**40.2**	13.5/37.9
ibo_Latn	**26.0**/-/**51.3**	23.9/50.4	**21.9**/-/**48.0**	**21.9**/46.1
lug_Latn	15.7/-/46.9	**25.8**/**55.2**	22.4/-/48.5	**30.9**/**54.4**
luo_Latn	12.0/-/39.4	**14.0**/**40.4**	14.3/-/38.3	**15.9**/**38.4**
swh_Latn	27.7/-/**57.2**	**30.7**/56.0	30.6/-/55.8	**39.3**/**60.8**
tsn_Latn	**31.9**/-/**59.5**	28.5/55.6	27.8/-/54.0	**37.3**/**60.2**
yor_Latn	13.9/-/**37.4**	**14.4**/36.3	18.0/-/41.0	**24.4**/**46.7**
zul_Latn	**22.9**/-/**56.3**	16.1/47.3	38.1/-/57.7	**40.3**/59.7
	fra-xx		xx-fra	
	Adelani et al. (2022)	NLLB-200	Adelani et al. (2022)	NLLB-200
bam_Latn	**24.7**/-/**49.9**	7.7/29.9	**25.8**/-/**49.0**	14.6/37.5
ewe_Latn	**8.9**/-/**37.5**	8.3/36.4	11.6/-/37.2	**19.4**/**42.6**
fon_Latn	**7.4**/-/**28.5**	3.4/21.8	**9.9**/-/**28.9**	8.9/28.7
mos_Latn	2.2/-/16.8	**5.4**/**27.6**	4.1/-/18.8	**6.1**/**23.5**
wol_Latn	**12.7**/-/**35.8**	9.1/29.9	**11.5**/-/**35.3**	9.5/30.2

Таблица 38: **Сравнение с лучшими достижениями на тестовом наборе MAFAND-MT.** Мы представляем оценки BLEU/spBLEU/chrF++ для NLLB-200 и лучшие BLEU и chrF++ из работы Adelani et al. (2022). Лучшие оценки в каждом направлении выделены жирным шрифтом. Низкоресурсные языки выделены подчеркиванием.

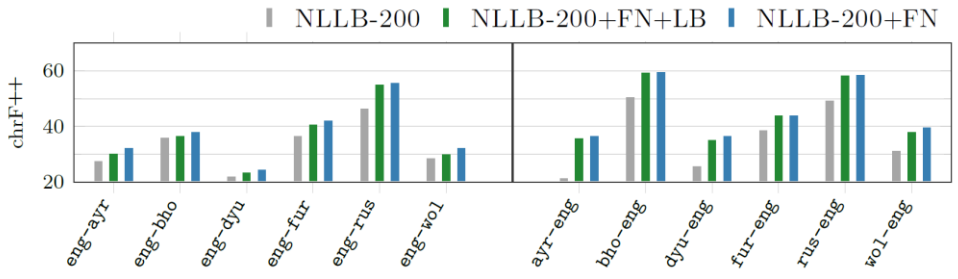


Рисунок 38: **Сравнение NLLB-200 с и без донастройки на 12 англоцентрических задачах NLLB-MD**. NLLB-200+FN+LB и +FN относятся к донастройке с и без балансировки нагрузки (LB). Мы представляем точность в терминах chrF++ на валидационном наборе.

Результаты. На рисунке 38 показаны оценки точности chrF++ на валидационных задачах домена чата для предобученной модели NLLB-200, модели, донастроенной с балансировкой нагрузки (NLLB-200+FN+LB), и модели, донастроенной без балансировки нагрузки (NLLB-200+FN).

В среднем донастройка (FN+LB) улучшает точность на +6,1 баллов по chrF++. Прирост производительности более значителен при переводе на языки с высокими ресурсами (eng и rus) на +8,9 баллов по chrF++ в среднем и на +2,0 баллов в среднем при переводе на 5 языков с низкими ресурсами в NLLB-MD. При отключении регуляризации балансировки нагрузки NLLB-200+FN улучшается на +7,2 балла по chrF++, что особенно интересно при переводе на языки с низкими ресурсами с приростом +3,7.

Затем мы донастраиваем с использованием нашей лучшей стратегии (NLLB-200+FN) в других трех областях NLLB-MD и представляем оценки chrF++ на тестовых наборах в Рисунке 39. В среднем, донастройкой NLLB-200 мы можем улучшить точность перевода в новых областях на +7,7 в чате, +3,1 в новостях, +4,1 в здравоохранении и +5,8 в скриптованных выступлениях (в терминах chrF++). Эти результаты свидетельствуют о трансферабельности и адаптивности NLLB-200 в другие области.

В прошлых работах возникла проблема донастройки разреженно активированных больших моделей (Artetxe et al., 2021; Fedus et al., 2022; Zoph et al., 2022). Эти большие модели более подвержены переобучению по сравнению с их плотными аналогами, и в некоторых случаях они показывали плохие результаты при донастройке (Artetxe et al., 2021; Fedus et al., 2022). Fedus et al. (2022) предлагает увеличить регуляризацию с помощью dropout экспертов, эффективно применяя более сильную регуляризацию к параметрам экспертов, в то время как Zoph et al. (2022) борются с переобучением, обновляя только подмножество параметров модели. С использованием маскировки вывода экспертов MoE (EOM), NLLB-200 сильно регуляризован и проявляет меньше переобучения на последующих задачах. Мы предполагаем, что без балансировки нагрузки мы позволяем модели отключать экспертов, фактически активируя только несколько, которые будут донастраиваться для последующей задачи. Это особенно важно при донастройке на одну задачу, для которой NLLB-200 научился назначать определенных экспертов (см. раздел 8.5.1); добавление потерь балансировки нагрузки, когда мини-пакеты не перемешиваются, значительно изменит эту наученную ассигновку. Мы оставляем исследование стратегий донастройки MoE с добавленной регуляризацией, выборочной донастройкой и расслабленной оптимизацией на будущее исследование.

8.5 Анализ NLLB-200

В этом разделе мы анализируем несколько свойств NLLB-200. Сначала мы обсуждаем языковую ко-локацию экспертов массово мультиязычной модели NLLB-200. Затем мы рассматриваем, как различные стратегии куррикулумного обучения решают проблему переобучения на направлениях с низкими ресурсами, сохраняя при этом производительность на направлениях с высокими ресурсами. Наконец, мы глубже вдаемся в изучение влияния мультиязычного трансфера на языки с низкими ресурсами.

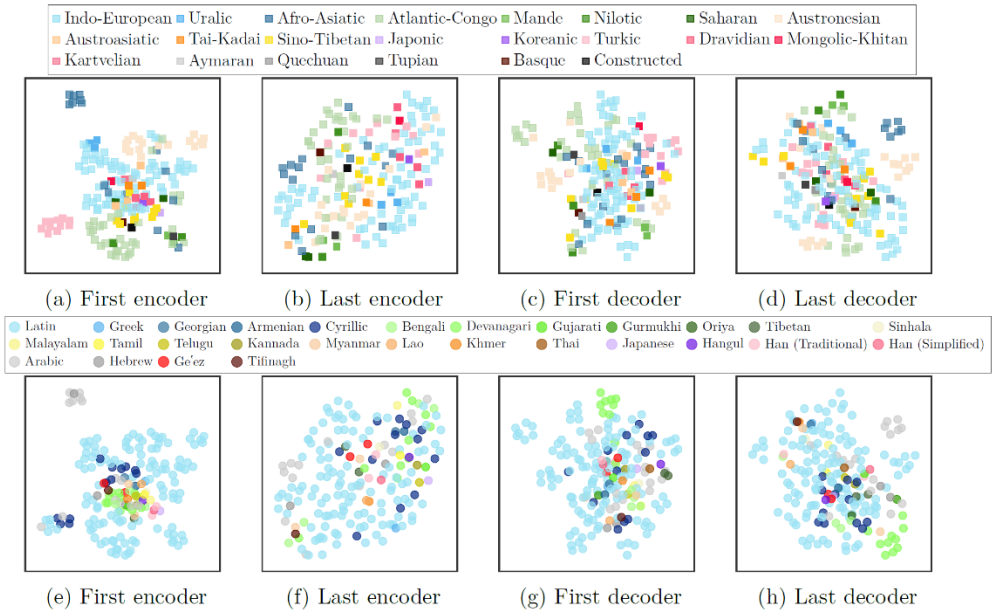


Рисунок 40: **Вложение UMAP для языков FLORES-200**. Мы вкладываем вектора назначения экспертов на язык в двумерное пространство с помощью UMAP. Языки с схожим выбором экспертов (усредненным на наборе данных dev Flores-200) расположены рядом в проекции на двумерное пространство. В первом ряду цвет отображен в соответствии с языковой семьей, а во втором ряду - в соответствии с алфавитом.

8.5.1 Ко-локация языков в экспертах NLLB-200

Аналогично нашему анализу моделей MoE на наборе данных с абляцией из 53 языков (см. раздел 6.2.4), мы стремимся проанализировать ко-локацию языков в экспертах NLLB-200. Следуя тем же шагам, мы вычисляем распределение для каждого языка среди 128 экспертов, оцениваем на наборе данных dev Flores-200, затем вкладываем все обученные языки в двумерное пространство с помощью UMAP. Аналогично нашим наблюдениям на наборе данных с абляцией, на Рисунке 40 мы видим, что поздние слои декодера и ранние слои энкодера NLLB-200 обрабатывают токены из разных языков, направляя их в разные наборы экспертов. Языки из одной семьи назначаются схожим наборам экспертов, а семьи, которые географически близки (например, нилотские, сахарские и атланто-конгские) или генеалогически связанные (например, арабский и иврит), также получают назначение к схожим экспертам MoE NLLB-200. Подробные оценки сходства между 200 языками в каждом из четырех изображенных слоев можно найти на Рисунке 45.

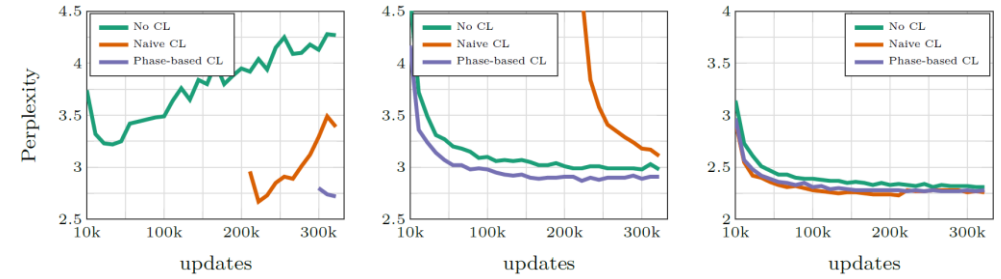


Рисунок 41: **Валидационная перплексия с различными стратегиями куррикулума**. Наша предложенная фазированная стратегия куррикулума особенно полезна для пар с низкими и очень низкими ресурсами, смягчая переобучение, наблюдаемое в базовых и наивных вариантах куррикулума. С наивным куррикулумом введение пар слишком рано (a) или слишком поздно (b) вредит производительности. Высокоресурсные пары практически не затрагиваются куррикулумом (c).

8.5.2 Влияние фазированного куррикулума на переобучение на языках с низкими ресурсами

На Рисунке 41 мы сравниваем эффект отсутствия куррикулума, наивного куррикулума и фазированного куррикулума при обучении NLLB-200. Мы замечаем, что наивный куррикулум помогает смягчить некоторое переобучение на языках с низкими ресурсами, однако некоторые такие пары вводятся слишком рано, а другие слишком поздно. Эти эффекты уменьшаются с использованием фазированного куррикулума. Эффекты более заметны для пар с очень низкими ресурсами, которые начинают значительно переобучаться, если их вводить слишком рано, как показано на Рисунке 41(a). Хотя стратегия фазированного куррикулума в целом помогает смягчить многие проблемы переобучения, наблюдаемые с наивным куррикулумом, она все еще не оптимальна. Разработка автоматизированных куррикулумов, которые мониторят паттерны переобучения в процессе обучения, может дальше улучшить производительность, и мы оставляем эту направление для будущих исследований.

8.5.3 Влияние мультиязычного трансфера

Одно из основных преимуществ мультиязычных моделей заключается в сильном трансфере между языками. Для количественной оценки выгод, полученных с помощью массово мультиязычной модели NLLB-200, мы обучили двуязычные модели на нескольких направлениях с низкими ресурсами с использованием тех же обучающих данных и сравнили их производительность. Результаты представлены в Таблице 39. Мы наблюдаем впечатляющие приросты почти во всех направлениях, на которых проводили сравнение, что демонстрирует выгоды мультиязычного трансфера. Также отмечаем, что приросты более последовательны для направлений xx-eng_Latn. Для языковых пар eng_Latn-xx улучшения менее последовательны, и мы можем наблюдать направления, где достигаются значительные приросты (+19,4 балла по chrF++), но также есть направления, где наблюдаются небольшие уменьшения (-0,6 балла по chrF++).

8.6 Сделать большие модели доступнее через дистилляцию

Хотя большие модели часто обладают высокой производительностью, их большие размеры ограничивают их полезность и делают выводы дорогостоящими. Добавление дополнительных языков, донастройка или даже генерация переводов могут потребовать значительных вычислительных ресурсов на GPU. В этом разделе мы исследуем дистилляцию, распространенную технику обучения более маленьких студенческих моделей на основе более крупных, более производительных учителей (Buciluǎ et al., 2006; Hinton et al., 2015; Tan et al., 2018). Такие техники обеспечивают более высокую производительность по сравнению с обучением небольшой модели с нуля (Gordon and Duh, 2020; Kim and Rush, 2016).

	eng_Latn-xx					xx-eng_Latn				
	Bilingual	Multilingual Dense		Multilingual MoE		Bilingual	Multilingual Dense		Multilingual MoE	
	chrF++	chrF++	Δ	chrF++	Δ	chrF++	chrF++	Δ	chrF++	Δ
aka_latn	16.8	35.6	**18.8**	36.2	**19.4**	36.6	45.6	**9.0**	46.7	**10.1**
smo_Latn	49.4	50.5	**1.1**	50.2	**0.8**	49.7	57.1	**7.4**	58.1	**8.4**
sot_Latn	44.4	46.4	**2.0**	46	**1.6**	52.3	61.9	**9.6**	62.5	**10.2**
twi_Latn	36.7	38.7	**2.0**	39.1	**2.4**	39.8	46.6	**6.8**	48	**8.2**
umb_Latn	23.3	23.8	**0.5**	24.1	**0.8**	27.2	31.6	**4.4**	32.7	**5.5**
vec_Latn	40.6	47.1	**6.5**	48.9	**8.3**	56.3	67.9	**11.6**	63.7	**7.4**
guj_Gujr	51.7	50.6	-1.1	51.4	-0.3	57.4	65.3	**7.9**	66.4	**9.0**
mya_Mymr	27.1	32.7	**5.6**	32.6	**5.5**	43.5	56	**12.5**	57.2	**13.7**
npi_Deva	32.4	47.9	**15.5**	48.6	**16.2**	56.2	67.3	**11.1**	67.6	**11.4**
pbt_Arab	37.1	35.3	-1.8	36.5	-0.6	47.2	56.6	**9.4**	56.9	**9.7**
sin_Sinh	41.8	43.1	**1.3**	44.1	**2.3**	51.0	61.7	**10.7**	63.0	**12.0**

Таблица 39: **Сравнение производительности FLORES-200 devtest между двуязычными и мультиязычными плотными и MoE моделями**. Мы видим, что большинство языков с низкими ресурсами получают выгоду от мультиязычного трансфера, что приводит к значительному увеличению баллов chrF++. Как плотные, так и MoE модели показывают лучшие результаты, и приросты более последовательны при переводе на eng_Latn.

Были исследованы различные подходы для сжатия моделей перевода (Freitag et al., 2017; Zhang et al., 2018; Zhou et al., 2020), включая мультиязычные (Sun et al., 2020) и условия с низкими ресурсами (Ahia et al., 2021; Saleh et al., 2020). Мы исследуем и обсуждаем два

типа подходов к дистилляции: онлайн и оффлайн дистилляция. Мы подробно описываем, как мы применяем эти техники для создания специализированных моделей для инструмента перевода контента Wikipedia (Laxström et al., 2015). Эти модели обеспечивают переводы на 25 языках и 74 направлениях с минимальной задержкой по требованию. Наконец, мы применяем эффективный подход для дистилляции нашей модели NLLB-200 с 54 миллиардами параметров на все 200 языков. Для практиков машинного обучения мы надеемся, что доступ к таким моделям позволит обеспечить переводы на значительно большее количество языков в различных приложениях.

8.6.1 Дистилляция знаний

В широком смысле, дистилляция знаний представляет собой процесс передачи знаний от более крупной учителеской модели к более маленькой студенческой модели (Hinton et al., 2015). Мы рассматриваем две формы дистилляции: онлайн и оффлайн. Кратко описываем оба этих подхода здесь.

Онлайн дистилляция. Мы исследуем дистилляцию знаний на уровне слов (Hinton et al., 2015). В этом случае студенческая модель обучается на обучающих данных, но с дополнительной целью: минимизировать перекрестную энтропию относительно распределения на уровне слов учителя.

Оффлайн дистилляция. Мы исследуем дистилляцию знаний на уровне последовательности (Kim and Rush, 2016), которую можно рассматривать как форму аугментации данных (Gordon and Duh, 2019; Xu et al., 2021b), аналогичную обратному переводу. Учительская модель используется для генерации переводов, используя монолингвальные данные в качестве входных данных. Затем студенческая модель обучается на данных, сгенерированных учителем. Этот подход позволяет студенческой модели научиться имитировать учителя на уровне предложений, но не на уровне предсказаний отдельных слов.

8.6.2 Создание моделей, специализированных для домена Wikipedia

В наших интервью с носителями языков с низкими ресурсами общей темой был доступ к информации онлайн, такой как образовательный контент. Хотя знания были отобраны на глобальном уровне на платформах вроде Wikipedia, эти платформы остаются доступными в основном для тех, кто говорит на одном из нескольких языков, доминирующих в интернете, таких как английский. Исходя из нашего принципа приоритета обслуживания недооцененных сообществ и обмена, мы следующим шагом исследуем, как наша работа может быть внедрена на практике. Мы описываем, как мы строим мультиязычные модели перевода для инструмента перевода контента Wikipedia. Это онлайн-интерфейс, который могут использовать редакторы Wikipedia для перевода статьи, которая уже существует на другом языке, в качестве отправной точки для написания новой статьи. Это способствует созданию новых статей на языках с недостаточным обслуживанием. Мы используем процесс дистилляции для создания моделей, которые достаточно быстры для удовлетворения требований к задержке услуги по запросу, сохраняя при этом высокое качество переводов. Такая работа имеет широкое применение для практиков машинного обучения, которые могут быть заинтересованы в адаптации универсальных систем под конкретные области и внедренные приложения.

Языковые пары. Языковые пары включают eng_Latn, fra_Latn, spa_Latn в качестве исходных языков, и 25 целевых языков, где языки либо (1) еще не поддерживаются в инструменте перевода контента Wikipedia, (2) поддерживаются, но производят переводы низкого качества, или (3) имеют высокие показатели удаления статей. Кроме того, мы включаем еще четыре других направления, интересных для Wikipedia: por_Latn-oci_Latn, cat_Latn-oci_Latn, zho_Hans-yue_Hant и rus_Cyrl-bak_Cyrl.

Обучение учителя и донастройка. Мы обучаем плотную учителескую модель с 1.3 миллиардами параметров на абляционном наборе данных (см. Раздел 6.4) для всех перечисленных языковых пар выше. Учителеская модель меньше нашей полной модели NLLB-200 с 54 миллиардами параметров и плотная, а не разреженная, что сокращает время, необходимое для создания обучающего набора данных студенческой модели для оффлайн дистилляции. Далее мы донастраиваем эту модель на битекстах домена Wikipedia, включая правки контента перевода Wikipedia. Мы фильтруем данные донастройки с использованием LID и фильтрации на основе длины, аналогичной описанной в разделе 8.1.4.

Применение оффлайн дистилляции. С использованием оффлайн дистилляции на уровне последовательности (Kim and Rush, 2016) мы дистиллируем ту же самую донастроенную учителескую модель в более маленькую плотную модель с 500 миллионами параметров. Мы подготавливаем обучающий набор данных для дистиллированной модели, генерируя переводы для монолингвальных корпусов данных Wikipedia с использованием донастроенной учителеской модели. Чтобы сгенерировать этот обучающий набор данных для студенческой модели, мы запускаем поиск лучших вариантов с размером пучка 4 с учителем модели. Исходные предложения берутся из дампов данных монолингвальных Wikipedia для eng_Latn, fra_Latn, spa_Latn, rus_Cyrl, por_Latn, cat_Latn. Наконец, мы обучаем студенческую модель на этом дистиллированном обучающем наборе данных. Обратите внимание, что в то время как учительская модель обучена в обе стороны, дистиллированная студенческая модель обучается только на направлениях eng_Latn-xx, spa_Latn-xx, fra_Latn-xx, por_Latn-oci_Latn, cat_Latn-oci_Latn, zho_Hans-yue_Hant и rus_Cyrl-bak_Cyrl.

Применение онлайн дистилляции. Мы используем онлайн дистилляцию на уровне слов (Hinton et al., 2015), чтобы дистиллировать донастроенную учителескую модель с 1.3 миллиарда параметров в более маленькую плотную модель с 500 миллионами параметров для повышения эффективности вывода. Мы используем те же монолингвальные дампы данных Wikipedia в качестве исходных предложений и применяем только мягкую кросс-энтропию относительно распределения учителя LKD.

Size	teacher 1.3B	teacher ft 1.3B	offline student 500M	online student 500M
eng_Latn-xx	40.0	43.0	**43.4**	43.0
fra_Latn-xx	36.3	39.2	**39.6**	39.1
spa_Latn-xx	35.1	36.9	**37.3**	36.8

Таблица 40: **Результаты оффлайн и онлайн дистилляции в настройке Wikipedia**. Мы показываем среднюю производительность chrf++ на тесте Flores-200 для английского, французского и испанского исходных языков для моделей учителя, донастроенной модели учителя и студенческой модели. Полный список результатов доступен в Таблице 59.

Сравнение онлайн и оффлайн дистилляции. Сначала мы рассмотрим производительность онлайн дистилляции. Из Таблицы 40 видно, что студенческая модель с 515 миллионами параметров показывает производительность сравнимую с донастроенной моделью учителя, отличаясь не более чем на 0.1 chrf++ для направлений eng_Latn-xx, fra_Latn-xx и spa_Latn-xx.

Для оффлайн дистилляции из Таблицы 40 видно, что студенческая модель с 515 миллионами параметров показывает лучшие результаты по сравнению с донастроенной моделью учителя с 1.3 миллиарда параметров в среднем по всем парам направлений. Наблюдаются улучшения для направлений eng_Latn-xx (+0.4 chrf++), fra_Latn-xx (+0.4 chrf++) и spa_Latn-xx (+0.3 chrf++). Мы предполагаем, что эти улучшения частично обусловлены использованием монолингвальных данных Wikipedia в качестве обучающего набора данных для дистилляции. Кроме того, дистиллированные модели часто производят больше транслейтов, что может повлиять на надежность метрик на основе ссылок (Riley et al., 2020).

Исходя из этих результатов, мы заключаем, что оффлайн дистилляция на уровне последовательностей выполняется немного лучше, чем онлайн дистилляция на уровне слов при дистилляции наших моделей для перевода контента Wikipedia. Однако онлайн дистилляция все равно демонстрирует хорошие результаты и является более эффективной с вычислительной точки зрения, поскольку для авторегрессивной генерации миллионов переводов не требуется очень большая MoE модель, что важно при дистилляции более крупных моделей на большее количество языков. В следующих разделах мы исследуем вычислительно эффективную онлайн дистилляцию для масштабных многоязычных моделей, таких как NLLB-200.

Сравнение с Bapna et al. (2022) Хотя значительные различия не позволяют справедливо сравнивать, мы кратко сравниваем нашу дистиллированную модель для Wikipedia с дистиллированными моделями Bapna et al. (2022) на Flores-101 devtest. Помимо основного различия в данных для обучения как для учителя, так и для студенческих моделей, существуют несколько других заметных отличий. Наша модель начинается с 1.3 миллиарда параметров плотного учителя, а не с модели на 6 миллиардов параметров. Мы охватываем 25 целевых языков по направлениям исходных языков eng_Latn, fra_Latn, spa_Latn и 79 направлений в одной студенческой модели, вместо двух студенческих моделей eng_Latn-xx и xx-eng_Latn, каждая поддерживающая 30 языков и направлений. Наша окончательная студенческая модель составляет всего 500 миллионов параметров, по сравнению с 850 миллионами параметров модели кодировщика с трансформером и декодером LSTM, представленной в работе (Bapna et al., 2022). Несмотря на это, при усреднении производительности по 6 пересекающимся направлениям на Flores-101 (eng_Latn-xx для asm_Beng, ckb_Arab, lin_Latn, lug_Latn, nso_Latn и gaz_Latn), наша дистиллированная модель с 500 миллионами параметров показывает схожую производительность с моделью с 850 миллионами параметров, дистиллированной в работе (Bapna et al., 2022). По 6 пересекающимся направлениям обе модели достигают 24.9 spBLEU на Flores-101 devtest.

8.6.3 Дистилляция NLLB-200, модели MoE с 54 миллиардами параметров

Мы окончательно исследуем дистилляцию NLLB-200, модели Mixture-of-Experts с 54 миллиардами параметров, с помощью онлайн-дистилляции. Окончательная дистиллированная модель сохраняет полную поддержку перевода всех 202 языков. Поскольку вывод для полной модели NLLB-200 медленнее, чем для плотной модели с 1.3 миллиарда параметров, и мы дистиллируем более чем в 30 раз больше пар языков, чем модель для Wikipedia, мы выбираем использовать онлайн-дистилляцию, а не оффлайн-дистилляцию из-за времени и вычислительных ресурсов, необходимых для генерации выходов учителя на уровне последовательности. Мы тренируем базовые и дистиллированные плотные модели, каждая с 1.3 миллиарда и 615 миллионами параметров, на протяжении 200,000 обновлений.

Результаты. В таблице 41 показано, что в среднем дистиллированная плотная модель студента с 1.3 миллиарда параметров показывает лучшие результаты по сравнению с базовой плотной моделью с 1.3 миллиарда параметров на +0.5 chrF++, а модель студента с 615 миллионами параметров лучше, чем базовая модель с 615 миллионами параметров на +0.3 chrF++. Мы замечаем, что дистилляция обеспечивает более значительные улучшения по сравнению с соответствующими базовыми плотными моделями для языков с низким и очень низким уровнем ресурсов. Например, для настройки модели с 1.3 миллиарда параметров дистиллированная модель показывает лучшие результаты по сравнению с плотной базовой моделью на +0.6 chrF++. Однако в среднем разрыв составляет -1.4 chrF++ для модели с 1.3 миллиарда и -3.7 chrF++ для модели с 615 миллионами между моделью студента и производительностью NLLB-200.

8.6.4 Заключение

Практическое внедрение моделей машинного обучения, особенно нейронных методов, является крайне сложным и активно изучаемым направлением исследований. Наше исследование показывает, что дистилляция является очень перспективным направлением для использования мультиязычных моделей и их специализации на подмножестве желаемых языковых направлений и в желаемой области. Это позволяет создавать модели перевода для Wikipedia с сильной производительностью (несмотря на то, что учительская модель имеет всего 1.3 миллиарда параметров). Мы надеемся, что такая работа и открытие доступа к дистиллированным моделям позволят другим получить переводы на свои родные языки.

8.7 Эффективное включение языков с несколькими письменностями и родственными лингвоидами

Языки не являются монолитными единицами, а являются изменчивыми и полными вариаций. Некоторые языки естественно пишутся на нескольких письменностях — например, сербский язык, который использует как латинскую, так и кириллическую письменности. Другие языки имеют большое количество вариаций — их мы называем лингвоидами. Эта вариация может быть относительно небольшими различиями в правописании (например, британский и американский английский) до такого уровня, что языки становятся взаимно несовместимыми.

В Flores-200 (см. раздел 3) мы включаем несколько письменностей для языков и несколько арабских лингвоидов. В данном разделе мы изучаем, как наилучшим образом представлять естественную вариативность языков и создавать наиболее точные локализованные переводы. Мы фокусируемся на двух случаях: транслитерации между различными письменностями и переводе близких арабских лингвоидов.

8.7.1 Транслитерация

В мире существует множество различных систем письма. Примеры включают логосиллабарии, такие как китайская письменность Хань, использующая идеографы для представления слов или морфем; слоговые письменности, такие как катакана, представляющие слоги; абугиды, такие как деванагари, в основе которых лежат пары согласных и гласных; абджады, например, арабская письменность, которая требует только записи согласных и может разрешать представление гласных в виде диакритических знаков; и алфавиты, такие как латинская и кириллическая системы письма, символы которых обозначают как гласные, так и согласные; и многие другие. В многих ситуациях один и тот же язык может использоваться через разные системы письма. Такие случаи возникают из-за исторических, геополитических, религиозных или технологических причин. Границы языков редко точно совпадают с границами геополитических сущностей, таких как национальные государства. Существование нескольких систем письма для одного языка приводит к двум основным проблемам для создателей языковых технологий, особенно когда речь идет о языках с низкими ресурсами или о языках, находящихся под угрозой исчезновения.

Проблемы с данными. Получение моделей транслитерации может быть сложной задачей. Независимо от того, является ли тестируемая техника правиловой или модельной, количество готовых к использованию данных может быть недостаточным. Переход от абджада к алфавиту с использованием правиловой техники, выведенной индуктивно из таблиц сопоставления символов, где особенно сложно описать сопоставление диакритических знаков, не приводит к удовлетворительным результатам. Аналогично модельные методы требуют достаточно больших объемов текстовых ресурсов (Madhani et al., 2022), которые по определению недоступны для языков с низкими ресурсами. Разделение лингвистических данных по разным системам письма может дополнительно снизить количество текста, которое можно использовать для обучения, что приводит к низкой производительности.

Этические соображения. Разработка технологий, которые предпочитают одну систему письма, может ненамеренно усиливать маргинализацию меньшинств, использующих другую систему письма. В данном разделе мы исследуем вопрос о том, может ли быть разработано технологическое решение, которое сможет смягчить эти опасения. Насколько эффективны наши модели перевода в генерации и ассимиляции контента в любой системе письма? Могут ли они использоваться для транслитерации с целью устранения разрывов между системами письма, или традиционные методы транслитерации эффективнее?

Экспериментальные настройки. Для оценки мы выбираем сложную задачу транслитерации из абджада в алфавит, что требует восстановления гласных. Мы оцениваем транслитерацию из арабской письменности в латинскую для языков ачехский и банджар, а также из тифинага в латиницу для языка тамашек. Системы, которые мы сравниваем:

	ace_Arab-ace_Latn	bjn_Arab-bjn_Latn	taq_Tfng-taq_Latn
uroman	0.47	0.40	0.23
Ejawi/ALA-LC	0.66	0.72	0.24
Ejawi/ALA-LC + rules	0.37	0.35	**0.22**
Custom Model	0.32	0.31	0.29
Rules + Custom Model	**0.25**	**0.20**	**0.22**
NLLB-200 translation model	0.54	0.41	0.66

Таблица 42: **Результаты транслитерации различными подходами**, измеренные по коэффициенту ошибок символов (CER, меньше - лучше) на тестовом наборе Flores-200.

• Универсальный инструмент универсальной романизации uroman (Hermjakob et al., 2018)

• Два онлайн-инструмента транслитерации: Ejawi43 и ALA-LC транслитератор для письменности Tifinagh script44

• Наш собственный переводческий модель NLLB-200, способная переводить между любой парой из 200 языков

• Настраиваемая нейронная модель транслитерации. Модель обучена на NLLB-Seed и основана на стандартной архитектуре трансформера с 4-слойным кодировщиком и декодером, 4 головками внимания и векторными представлениями размерности 256. Размер словаря SentencePiece для каждого языка определяется как функция размера набора символов. Мы экспериментируем с размерами словаря в 1.5, 2.5 и 10 раз больше размера набора символов и выбираем лучший вариант для каждого языка через валидацию. Подобные архитектуры уже показали свою эффективность в задачах трансдукции символов и нормализации текста с низкими ресурсами (Lusito et al., 2022; Wu et al., 2021).

• Правила, основанные на подходе, который наивно заменяет отдельные символы согласно таблице транслитерации. Мы комбинируем этот подход с Ejawi в качестве шага постобработки, поскольку заметили, что некоторые слова остаются в исходной письменности после использования этого онлайн-инструмента. Мы также пытаемся использовать этот подход на этапе предварительной обработки для нейронного транслитератора с целью приближения векторных представлений исходного и целевого текста друг к другу.

Оценка. Мы используем CER (коэффициент ошибок символов), который вычисляется как соотношение редактированных символов к эталонным символам, что является распространенным подходом в задачах, таких как оптическое распознавание символов. Все системы оцениваются на наборе данных Flores-200 devtest.

Результаты. Результаты представлены в Таблице 42. Мы видим, что специализированные модели транслитерации лучше всего справляются с этой задачей. Среди этих инструментов мы отмечаем, что онлайн-сервис транслитерации Ejawi демонстрирует высокий уровень ошибок сам по себе, но значительно улучшает результаты, если за ним следует простой этап замены символов на основе правил.45 Нейронные системы транслитерации, которые мы тренировали, в целом не достигают самых низких уровней ошибок, но их производительность, вероятно, можно улучшить путем сбора больших обучающих наборов данных. Вне зависимости от подхода, уровни ошибок остаются относительно высокими. Это подчеркивает важность тщательного подхода при разработке языковых технологий, чтобы минимизировать различия, которые могут возникнуть для пользователей различных письменностей.

8.7.2 Мультидиалектальный перевод

Мы хотим рассмотреть возможность перевода на или с различных арабских лингвистических форм. Для тестирования возможности и эффективности мультиязычных моделей перевода для этой задачи мы здесь фокусируемся на арабских лингвистических формах, представленных в Таблице 43. Арабский язык имеет преимущество в богатстве диалектного разнообразия, которое было хорошо изучено и задокументировано (Alshargi et al., 2019; Habash, 2010; Habash et al., 2013; Salameh et al., 2018).

Арабские Лингвистические Формы. В то время как системы обработки естественного языка исторически сосредотачивались на современном стандартном арабском языке (ССА), в последние годы было проведено значительное количество исследований по переводу арабских лингвистических форм (Baniata et al., 2021; Kumar et al., 2021; Sajjad et al., 2020). NLLB-200 включает несколько других разновидностей арабского языка помимо ССА: египетский, марокканский, наджийский, северо-левантийский, южно-левантийский, таиззи-аденский, месопотамский и тунисский. Этот экспансионный подход требует четкого способа оценки различий между этими лингвистическими формами и ССА. Следует отметить, что общепризнанным характеристикой арабского является социолингвистическое явление, известное как диглоссия. В диглоссических ситуациях пользователи переходят от более разговорного диалекта, используемого для неформальных и повседневных задач (например, разговоры, чаты или посты в социальных сетях), к другому диалекту, используемому для задач, которые обычно выполняются на более высоком уровне языка (например, технический или научный дискурс, образовательный или информационный контент; см. Ferguson, 1959). Такова ситуация с арабскими лингвистическими формами, где ССА используется для задач более высокого уровня. Прямым следствием этого явления является то, что предложения, созданные на любой арабской лингвистической форме на более высоком уровне языка, вероятно, будут иметь больше общего с ССА, чем предложения, созданные на более низком уровне (например, неформальные чаты или посты в социальных сетях).

Уровень диалектности и его расчет на наборе данных Flores-200. Уровень диалектности (DL; см. Sajjad et al., 2020) служит способом измерения степени лексической вариации в этих лингвистических формах по сравнению с ССА. Конкретно, DL представляет собой отображение количества токенов, присутствующих в лингвистической форме и отсутствующих в ССА, включая случаи семантических различий для токена. Это означает, что в случаях, когда значение данного токена в ССА отличается от значения того же токена в другой лингвистической форме, он не учитывается в оценке DL. DL представлен в диапазоне от 0% до 100%, где чем выше диапазон, тем выше уровень диалектности в конкретном диалекте.

DL вычисляется двумя способами для наших целей: на уровне корпуса и на уровне предложения. Первая метрика, DL на уровне корпуса (далее cDL), вычисляется для заданной не-ССА лингвистической формы как доля всех токенов в ее наборе данных Flores-200 devtest, которые отсутствуют в ССА. Второе измерение DL, DL на уровне предложения (далее sDL), вычисляется аналогично, но на более высоком уровне гранулярности. Мы определяем, для каждой строки в наборе данных Flores-200 devtest лингвистической формы, долю ее токенов, которые отсутствуют в соответствующем предложении на ССА. Это число затем усредняется по всем строкам в наборе данных. Как и в работе Sajjad et al. (2020), для вычисления этих метрик мы нормализуем наш арабский текст путем замены различных форм Алиф и Хамзы, а также использования индо-арабских цифр. Кроме того, мы также используем те же метрики, использованные во всей статье для оценки машинного перевода, такие как spBLEU и chrF++, описанные в разделе 7.

Результаты. Учитывая, что Flores-200 выровнен, мы начинаем с измерения различий между данными Flores-200 для шести арабских лингвистических форм: Месопотамского, Ta‘иззи-Адени, Тунисского, Южного Левантийского, Надждского и Египетского. Следует отметить, что данные для Южного Левантийского (ajp_Arab) и Египетского (arz_Arab) были получены путем перевода с английского, в то время как наборы данных других четырех лингвистических форм были адаптированы непосредственно из ССА. Хотя мы решаем включить эти две лингвистические формы в наши анализы для предоставления дополнительного контекста, следует иметь в виду, что результаты, касающиеся их, могут быть искажены из-за немного различного процесса сбора данных. По этой причине мы оставляем их отдельно в таблицах результатов. Результаты в Таблице 43 показывают, что за исключением двух лингвистических форм, данные для которых были получены через перевод, все остальные лингвистические формы близки к ССА, с показателями spBLEU, превышающими 50 баллов. Среди адаптированных наборов данных, тунисский (aeb_Arab) является наиболее отличающимся от ССА, а надждский (ars_Arab) практически полностью совпадает с ССА.

	Transl. perf.		against generated MSA			
Direction	spBLEU	chrF++	sDL	cDL	spBLEU	chrF++
fra_Latn-acm_Arab	3.7	18.4	65.80%	8.47%	7.3	25.4
fra_Latn-acq_Arab	19.1	34.3	33.06%	17.54%	37.1	49.4
fra_Latn-aeb_Arab	11.5	30.2	49.89%	10.49%	18.2	37.9
fra_Latn-ars_Arab	25.8	40.5	25.71%	11.16%	50.3	60.6
fra_Latn-ajp_Arab	27.0	42.8	29.41%	10.41%	49.0	59.2
fra_Latn-arz_Arab	25.3	41.2	29.86%	9.26%	49.8	62.5

Таблица 45: **Сравнение переводов, сгенерированных NLLB-200 с французского**. Данные devtest Flores-200, аналогично таблице 44.

	into eng_Latn		into fra_Latn	
Language	spBLEU	chrF++	spBLEU	chrF++
acm_Arab	43.3	63.1	39.2	57.6
acq_Arab	45.1	64.9	40.2	58.5
aeb_Arab	38.6	59.5	36.0	55.1
ars_Arab	46.8	66.0	41.6	59.5
ajp_Arab	48.2	67.4	41.8	59.9
arz_Arab	40.7	60.8	37.8	56.1
arb_Arab	48.3	66.9	42.3	59.8

Таблица 46: **Результаты работы языковых моделей на тестовом наборе данных FLORES-200**. Мы оцениваем перевод на английский и французский для семи арабских языковых моделей, включая MSA (низ списка).

Производительность NLLB-200. Далее мы переходим к тому, как NLLB-200 генерирует арабские языковые модели. Таблицы 44 и 45 рассматривают перевод модели с английского и французского соответственно. Первые два столбца измеряют производительность модели по переводу на каждую языковую модель по данным ссылки Flores-200. Мы видим, что среди языковых моделей, чьи данные оценивались на основе набора данных MSA (первые четыре строки), наблюдаемые тенденции из Таблицы 43 все еще заметны. Производительность низкая для тунисского (aeb_Arab) и месопотамского (acm_Arab), и лучшая для наджийского (ars_Arab), что, вероятно, объясняется тем, насколько каждая модель выигрывает от передачи знаний через значительно более ресурсные данные MSA. Напротив, два направления, данные оценки которых были непосредственно переведены с английского (нижние две строки), достигают относительно высоких показателей перевода, несмотря на то что их данные ссылки показывают наибольшие различия с MSA. Мы можем сформулировать две гипотезы по этому поводу. Во-первых, независимо от диалектальных вариаций, показанных для этих языков в Таблице 43, они, возможно, действительно близки к MSA, и их воспринимаемые диалектальные различия в значительной степени могут быть объяснены немного отличающейся процедурой, используемой для сбора данных оценки. Во-вторых, поскольку их данные оценки были непосредственно переведены с английского (в отличие от данных, адаптированных из MSA, который в свою очередь был переведен с английского), они могут быть ближе семантически и синтаксически к английскому источнику, что делает задачу их перевода более простой.

	Custom			NLLB-200		
Directions	spBLEU	chrF++	#best	spBLEU	chrF++	#best
{Arabic dialects}-eng_Latn	34.1 (4.7)	56.7 (3.7)	0	**44.4** (3.8)	**64.1** (3.1)	**7**
{Arabic dialects}-fra_Latn	19.9 (1.4)	41.0 (1.3)	0	**39.8** (2.3)	**58.1** (1.9)	**7**
eng_Latn-{Arabic dialects}	**32.0** (6.3)	**47.5** (5.7)	**4**	29.5 (10.8)	45.4 (8.6)	3
fra_Latn-{Arabic dialects}	**31.8** (5.7)	**47.1** (5.5)	**7**	21.1 (10.5)	36.7 (10.2)	0
Overall	29.4 (7.3)	48.1 (7.1)	11	**33.7** (11.9)	**51.1** (12.7)	**17**

Таблица 47: **Сравнение пользовательской арабской модели и NLLB-200** на тестовом наборе данных FLORES-200. Мы сообщаем среднюю производительность и стандартное отклонение для перевода арабских языковых моделей на и из eng_Latn и fra_Latn. Также мы сообщаем количество направлений в каждом наборе, для которых данная модель достигает лучшей производительности.

Затем мы анализируем, насколько различаются диалекты, генерируемые нашей моделью перевода. В самой правой части Таблиц 44 и 45 мы измеряем различия между сгенерированным арабским текстом, не относящимся к MSA, и сгенерированным MSA. Для первых четырех языковых моделей в группе основные тенденции, выявленные в Таблице 43, все еще действительны, с тунисским и месопотамским языками, наиболее отдаленными от MSA, и наджийским языком, наиболее близким к нему. Однако более общим образом мы наблюдаем небольшое снижение различий в уровнях диалектности текста, переведенного с английского, по сравнению с оценочными данными, аннотированными человеком. Действительно, оценочные баллы DL на уровне предложений имеют стандартное отклонение 14.6 для оценочных данных, которое снижается до 8.8 для данных, сгенерированных моделью.

Кроме того, мы изучаем способность переводящей модели обрабатывать различные арабские языковые модели, рассматривая ее производительность при их переводе на английский и французский языки. Результаты представлены в Таблице 46. Как и ожидалось, для обеих целевых языков наиболее высокую производительность показывает высокоресурсный MSA (arb_Arab), за которым следуют малоресурсные южно-левантийский (ajp_Arab) и наджийский (ars_Arab) языковые модели, которые выигрывают от передачи знаний. Среди четырех языковых моделей, данные оценки которых были собраны путем адаптации, мы видим, что относительный рейтинг производительности перевода соответствует порядку ранжирования уровня диалектности в Таблице 43: модель наилучшим образом понимает диалекты, наиболее близкие к MSA.

Наконец, в Таблице 47 мы сравниваем производительность большой модели NLLB-200 с более целевой, меньшей моделью, которая фокусируется только на переводе между арабскими языковыми моделями, eng_Latn и fra_Latn. Эта пользовательская модель использует более компактную трансформаторную архитектуру с 12 слоями кодировщика и 12 слоями декодировщика, размером FFN 4096, 16 головками внимания и в остальном следует настройкам базовой модели, использованной в разделе 6.2.1. Мы видим, что в среднем NLLB-200 превосходит пользовательскую модель, достигая лучшей производительности в 17 из 28 направлений, а также имеет более высокий средний балл. Производительность большой модели особенно выше при переводе на французский язык. Это можно объяснить различием в количестве первичных тренировочных предложений fra_Latn, к которым обращается каждая модель. Модель NLLB-200, имеющая доступ к гораздо большему числу направлений, может оперировать более чем 53 миллионами уникальных первичных французских предложений. Ограничение корпусов обучения только англо-центричными и арабо-центричными направлениями уменьшает это число до менее чем 38 миллионов для пользовательской модели. С другой стороны, меньшая модель превосходит NLLB-200 при переводе на арабские языковые модели, показывая, что большая мультиязычная модель не всегда является равномерно лучшим выбором для данного набора языков.

Заключение. В заключение, эти результаты отражают другие подобные усилия, такие как бенчмарк AraBench (Саджад и др., 2020), подчеркивающий преимущества переводов для арабских языков, созданных с использованием системы MSA. Их выводы, связанные с уровнем диалектности и качеством перевода поколений, также можно увидеть в наших результатах. Как указано ранее, из арабских языков, на которые мы сосредоточились, наилучшие результаты показывает Нэджди, так как он имеет постоянно низкий уровень DL, что ведет к высокому качеству перевода. В отличие от этого, Месопотамский и Тунисский, показывающие высокий DL в Таблице 43, достигают постоянно низкой производительности перевода. Эти результаты, а также результаты бенчмарка AraBench, где заключают, что для переводов, выполненных с использованием системы MSA, язык с более низким DL будет порождать более качественный перевод, предоставляют способ преодоления ресурсного разрыва, который может существовать для этих языков. Их короткое лингвистическое расстояние до MSA позволяет им значительно выигрывать благодаря трансферному обучению, служащему средством улучшения машинного перевода с ограниченными ресурсами и мультидиалектального перевода. Наконец, наше сравнение NLLB-200 и более целевой модели показывает, что хотя многоязычная модель достигает лучшего среднего результата, более маленькая модель все еще может превосходить ее в определенных направлениях, подчеркивая важность более фокусированных исследований по близкородственным языкам.

8.8 Экологическое влияние NLLB

Оценки выбросов углерода не являются точными, поскольку у сообщества нет инструментов для точного измерения факторов, влияющих на выбросы. В предыдущих работах были представлены оценки и рекомендации в работах Бендера и др. (2021); Додж и др. (2022); Паттерсон и др. (2021); Уу и др. (2022). В данной работе мы полагаемся на лучшие доступные оценки потребления энергии GPU и углеродной эффективности. Следует отметить, что оценки облачных провайдеров все еще неточны. Есть несколько факторов, влияющих на точность этих измерений: реальное потребление энергии GPU зависит от его использования и, вероятно, отличается от теплового проектного параметра (TDP), который мы используем как потребление энергии GPU. Кроме того, мы не включили дополнительные затраты на энергопотребление, такие как потребление энергии InfiniBand (IB) или не-GPU потребление энергии серверов или охлаждение дата-центров. Кроме того, производственные углеродные затраты для AI систем, таких как GPU, могут внести дополнительный углеродный след (Гупта и др., 2022a,b). Мы надеемся, что анализ углеродного следа для NLLB поможет обеспечить прозрачность для понимания экологических последствий технологий AI.

Выбросы углерода при обучении NLLB 200. Обучение NLLB-200 производилось на GPU NVIDIA A100. Используя спецификации системы NVIDIA A100 (Шокет и др., 2021), мы используем TDP 400W как мощность на процессор. Для обучения NLLB-200 было проведено кумулятивно 51968 GPU-часов вычислений на оборудовании типа A100-SXM-80GB (TDP 400W). Мы оцениваем общие выбросы углерода при обучении NLLB-200 в размере 8.39 т CO2eq, из которых 100% были непосредственно компенсированы программой устойчивости поставщика.

	Time (h)	Power Consumption (W)	Carbon Emitted (tCO_2eq)
Data Mining	108,366	400	17.55
Backtranslation	18,000	300	2.17
Modeling	196,608	400	31.74
Final Ablations	224,000	400	36.17
Evaluations	51,200	400	8.26
NLLB-200	51,968	400	8.39
Total			104.31

Таблица 48: **Общий углеродный след для программы "No Language Left Behind".** Мы предоставляем подробные оценки для всех этапов, которые используют GPU для вычислений. Здесь мы перечисляем факторы, которые вошли в расчет: Время в часах - общее время работы GPU, необходимое для выполнения этапа. Потребление энергии - потребление энергии на одно устройство GPU для используемых GPU, скорректированное с учетом эффективности использования энергии. 100% выбросов напрямую компенсируются поставщиками.

Общий углеродный след всей программы "No Language Left Behind". Приведенные выше данные отражают только углеродный след нашей финальной модели. Однако существуют несколько этапов исследовательского процесса до обучения финальной модели и шаги после (такие как производство переводов и этапы оценки человеком), которые также необходимо учитывать (Wu et al., 2022). Большинство предыдущих работ просто сообщают углеродный след от обучения их крупнейших моделей и умножают его на коэффициент (обычно 2x), чтобы указать общие выбросы. Вместо этого мы пытаемся сообщить углеродный след для всех этапов, которые используют GPU. Это также предоставляет сообществу полезную информацию о вычислительных требованиях и эффективности каждого этапа. Наш подробный отчет включает этапы подготовки данных, включающие крупномасштабный майнинг битекстов (раздел 5.3) и обратный перевод (раздел 8.1.3), все модельные эксперименты по разработке нашей архитектуры и методов обучения (раздел 6), финальные абляции модели для всех 200 языков (раздел 8.2) и оценки моделей (раздел 8.3). В Таблице 48 мы подробно описываем каждый расчет и наблюдаем, что экспериментальная фаза нашего исследования является наиболее вычислительно затратной из всех. Общие выбросы для проекта NLLB в целом оцениваются в **104.31 т CO2eq, из которых 100% были непосредственно компенсированы поставщиком**.

Разреженные модели типа "смесь экспертов" могут иметь огромное количество параметров, но они активируются лишь частично при обработке токенов во время обучения. Следовательно, они могут быть более вычислительно эффективными по сравнению с их плотными аналогами, и масштабирование таких моделей с тщательным учетом может помочь снизить потребление энергии и выбросы углерода для больших нагрузок. Для исследователей машинного обучения крайне важно тщательно измерять и сообщать влияние их работы на выбросы углерода, оптимизировать свои рабочие нагрузки для энергоэффективных архитектур и парадигм обучения, а также помнить о негативном воздействии на окружающую среду неэффективных рабочих нагрузок. Большие рабочие нагрузки в области машинного обучения часто несут затраты для окружающей среды, и если такие работы не будут открыты для сообщества, аналогичные усилия придется дублировать в различных исследовательских группах, что приведет к дополнительным выбросам углерода. Чтобы помочь сократить необходимость дублирования аналогичных рабочих нагрузок, мы открываем исходный код NLLB-

200 и других меньших плотных моделей, дистиллированных моделей, нашего оптимизированного кода для обучения/инференса, результаты оценок и бенчмарки, а также метаданные для добытых битекстов. Все материалы и код доступны по ссылке:

9. Программа "No Language Left Behind": социальное воздействие и заключительные мысли

В рамках этого проекта мы взяли на себя задачу создания высококачественных систем машинного перевода для более чем 200 языков. Столкнувшись с такими серьезными препятствиями, как отсутствие надежных данных для оценки и обучения, прогресс в области перевода с ограниченными ресурсами идет медленнее по сравнению с переводом с высокими ресурсами. В NLLB мы используем новаторские подходы, чтобы сделать несколько значительных вкладов, направленных на устранение этих разрывов: (1) Flores-200, высококачественный набор данных для оценки, переведенный человеком, и NLLB-Seed, набор данных, включающий битексты, переведенные человеком, для 43 языков, (2) новаторский метод майнинга битекстов, создающий сотни миллионов согласованных тренировочных предложений для языков с ограниченными ресурсами с использованием нашей библиотеки майнинга с открытым исходным кодом Stopes и модели идентификации языка, и (3) различные методики моделирования, специально разработанные для значительного улучшения многоязычного перевода с ограниченными ресурсами за счет снижения переобучения. Кроме того, мы также создали меньшие, дистиллированные модели, чтобы исследовательское сообщество и различные специалисты по машинному обучению могли более легко внедрять эти разработки.

В заключение, мы обсудим потенциальное социальное воздействие нашей работы. Как и в случае с большинством достижений в области искусственного интеллекта, измерение социального воздействия NLLB требует систематической оценки и долгосрочного подхода. Хотя его внедрение может принести пользу нескольким заинтересованным сторонам, включая группы с ограниченными ресурсами и научное сообщество в целом, мы также признаем, что такое вмешательство может иметь потенциальные недостатки. Таким образом, мы рассматриваем возможности и ограничения NLLB и способы максимизации его преимуществ при минимизации вреда.

9.1 Расширение доступа к информации

Летом 2016 года ООН объявила доступ к интернету основным правом человека (Howell и West, 2016). Хотя целью этого заявления было побудить страны ограничить цензуру и позволить информации и идеям свободно циркулировать, большая часть интернета остается недоступной для многих из-за языковых барьеров. NLLB имеет потенциал изменить статус-кво, сделав интернет более доступным для многих.

Для многих языковых сообществ с ограниченными ресурсами предложение NLLB станет первой моделью, разработанной для поддержки перевода их языков. Пользователи инструментов NLLB смогут получить доступ к контенту, который ранее был для них недоступен, что позволит усилить их ознакомление с информацией и медиа. Хотя его воздействие может охватить многие области повседневной жизни, его влияние на образование, которое также изучалось в других исследованиях машинного перевода (Lee, 2020), может быть значительным. В формальных образовательных учреждениях, например, студенты и преподаватели, принадлежащие к группам с ограниченными языковыми ресурсами, смогут получить доступ к большему количеству книг, научных статей и архивов, чем раньше. В рамках неформального обучения носители языков с ограниченными ресурсами смогут получить больший доступ к информации от мировых новостных агентств и социальных медиа, а также к онлайн-энциклопедиям, таким как Wikipedia. В этих последних пространствах, где производство знаний и контента идет с головокружительной скоростью, значение перевода невозможно недооценить (Bywood et al., 2017; Singh et al., 2012).

Преимущества качественного перевода не ограничиваются только слабо обслуживаемыми сообществами. Для сообществ, которые в настоящее время обслуживаются другими сервисами перевода, улучшение качества перевода повысит их общую доступность и использование возможностей интернета. Такие улучшения качества также могут привести к более упорядоченному процессу приобретения знаний и коммуникации. Когнитивная энергия, сэкономленная на расшифровке плохо переведенного контента, может быть направлена на выполнение других более важных задач.

Поскольку язык неразрывно связан с культурой, для многих языков с ограниченными ресурсами, находящихся под угрозой исчезновения, угроза утраты языка также может означать разрушение наследия (Sallabank, 2013). NLLB может мотивировать большее количество писателей или создателей контента на языках с ограниченными ресурсами делиться локализованными знаниями или различными аспектами своей культуры как с культурными инсайдерами, так и с аутсайдерами через платформы социальных сетей или такие сайты, как Wikipedia. Предоставление людям доступа к новым инструментам перевода может таким образом открыть ценное направление для двустороннего обучения. В долгосрочной перспективе такие процессы создания контента могут повлиять на глобальную систему знаний, бросить вызов западоцентричным способам производства и распространения знаний и помочь в возрождении определенных меньшинственных культур и языков (Bird, 2019; Bird и Chiang, 2012).

9.2 Двуличная природа цифрового участия

Преимущества технологического вмешательства, такого как NLLB, необходимо тщательно взвешивать с учетом затрат и рисков, которые оно может принести группам, использующим языки с ограниченными ресурсами, и другим заинтересованным сторонам. Например, увеличение цифрового участия и языкового представительства повышает видимость группы (Bucher, 2012). Такая видимость может увеличить вероятность того, что группы станут объектами наблюдения и цензуры (Treré, 2016; Zuboff, 2019). Кроме того, затронутые сообщества могут стать более уязвимыми к дезинформации, онлайн-мошенничеству или разжиганию ненависти (Gereme et al., 2021; Hossain et al., 2020). Иными словами, расширение доступа к языкам и информации делает определенные группы более уязвимыми к давним проблемам, преследующим цифровые сообщества в целом. Хотя для этих сложных вопросов не существует простого решения, мы надеемся, что NLLB можно будет использовать для его кросс-лингвального потенциала (Conneau et al., 2018) для усиления существующих (и обычно одноязычных) инструментов, предназначенных для обнаружения и классификации разжигания ненависти, фишинга и других социально вредоносных онлайн-текстов на языках с ограниченными ресурсами (Khonji et al., 2013; MacAvaney et al., 2019). Понимая, что такие инструменты действуют как первая линия обороны, мы считаем, что долгосрочные структурные инвестиции, направленные на борьбу с вредоносной цифровой деятельностью и повышение онлайн-грамотности, должны идти рука об руку с внедрением новых инструментов, таких как NLLB.

Хотя доступ к переводу может повысить общее цифровое участие, он также может усугубить существующее цифровое неравенство на местном или общинном уровне. Например, те, кто владеет технологическими навыками, получат больше пользы от NLLB, чем те, кто ими не владеет. Демографически эти различия отражаются через такие факторы, как возраст, образование, экономическое положение и сельская местность (Elena-Bucea et al., 2021; Hindman, 2000). Более того, поскольку технологическая инфраструктура неравномерно распределена во многих частях мира, сообщества, которые уже отстают в доступе к интернету, могут столкнуться с еще большими информационными разрывами по сравнению с их лучше обслуживаемыми аналогами. Иными словами, неравенство в доступе к знаниям, социальным связям и экономической мобильности может углубиться, если не будут приняты структурные меры для исправления существующих проблем, влияющих на сообщества, использующие языки с ограниченными ресурсами.

Учитывая, что основная цель NLLB — уменьшить языковое неравенство в глобальном контексте, в долгосрочной перспективе все больше и больше языков с ограниченными ресурсами будут включены в проект (или подобные проекты). На этом пути исследователи неизбежно столкнутся с растущим числом уязвимых сообществ, которые могут сопротивляться идее позволить технологическим организациям, с которыми у них мало связей, использовать их языки в своих целях (Coffey, 2021). В этой связи крайне важно быть рефлексивными в нашем подходе и уделять приоритетное внимание отношениям с местными учреждениями и членами сообщества, чтобы лучше понять их потребности и опасения при любых расширениях проекта. Эта мотивация дополнительно объясняет, почему мы разработали долгосрочные углубленные интервью и полевые исследования с носителями языков с ограниченными ресурсами, чтобы понять, как наше вмешательство может повлиять на их повседневную жизнь. Более того, сотрудничество с исследовательскими группами, которые уже имеют заинтересованность в данной теме, является необходимым для любого будущего успеха. Мы надеемся, что акцент на взаимных интересах и общих моральных ценностях будет способствовать объединению ресурсов и знаний, прокладывая путь к долгосрочному сотрудничеству среди различных заинтересованных сторон.

9.3 Будущее NLLB: Коллективная ответственность

Признавая, что решение языковых различий с помощью машинного перевода — это огромная задача, решение NLLB сделать наборы данных и модели общедоступными способствует инновациям через совместное производство и сотрудничество (Weber, 2004). Открытие наших наборов данных и моделей не только пропагандирует прозрачность в разработке технологий ИИ, но также предотвращает дублирование усилий и позволяет практикам машинного перевода сосредоточить свою энергию на выявлении пробелов и построении на основе нашей работы (Kogut и Metiu, 2001). Кроме того, мы активно разрабатываем механизмы, которые позволят нам поддерживать ученых по данным и исследователей, желающих использовать и адаптировать наши модели для своих нужд, и предоставлять необходимую помощь при необходимости. Мы осознаем, что открытие исходного кода не означает равный доступ; технологические и инфраструктурные барьеры для развертывания наших моделей и наборов данных остаются высокими для исследователей во многих частях мира. Поэтому, чтобы смягчить проблемы, связанные с неравномерным распределением вычислительных мощностей, мы планируем разрабатывать инструментарии и выдавать гранты малообеспеченным лабораториям для помощи в их научных исследованиях. Мы верим, что сотрудничество и систематическая долгосрочная документация позволят нам лучше оценить наше влияние на различные сообщества, вовлеченные в наш проект.

Более того, совместное использование NLLB с более широкой научной и исследовательской общиной позволит специалистам с разнообразными компетенциями внести свой вклад в развитие проекта. Во многих отношениях состав усилий NLLB подчеркивает центральное значение междисциплинарности в формировании нашего видения. Машинный перевод находится на пересечении технологического, культурного и социального развития и требует от ученых с разным образованием и точками зрения полного понимания всех аспектов (Kusters et al., 2020). Мы надеемся, что в будущем NLLB продолжит расширяться, включая ученых из областей, мало представленных в мире машинного перевода и ИИ, особенно тех, кто имеет гуманитарное и социальное образование. Более важно, мы надеемся, что команды, разрабатывающие такие инициативы, будут представлять широкий спектр рас, гендеров и культурных идентичностей, подобных тем сообществам, чьи жизни мы стремимся улучшить.

Наконец, мы хотим подчеркнуть, что преодоление препятствий, мешающих вебу быть действительно доступным для носителей всех языков, требует многогранного подхода. Как единое технологическое вмешательство, NLLB является лишь одной частью огромной головоломки. Политические меры, направленные на более фундаментальные вопросы, касающиеся образования, доступа в интернет и цифровой грамотности, являются необходимыми для искоренения структурной проблемы языковых различий. Мы привержены совместной работе с различными заинтересованными сторонами, продолжая наш путь к созданию технологий перевода, которые делают веб более доступным местом, независимо от того, на каком языке говорит человек.

10. Вклады

Мы описываем вклады каждого члена проекта "No Language Left Behind", сгруппированные по секциям и отсортированные в алфавитном порядке по фамилии. Каждый человек упоминается только один раз, хотя многие внесли вклад в несколько областей. Ни одно количество места не может полностью описать страсть и вклад каждого человека, участвовавшего в реализации этого проекта.

Данные

- **Bapi Akula** - разработка монолингвальной обработки данных от CommonCrawl до дедуплицированных и отфильтрованных предложений

- **Pierre Andrews** - ведущий инженер по данным, возглавил разработку стопов, добычу данных и монолингвальные очистительные процессы

- **Onur Çelebi** - идентификация языка (LID) для более чем 200 языков, открытый исходный код добытых данных, реализация и обсуждения для разработки LASER3 и монолингвального пайплайна

- **Kenneth Heafield** - техническая обратная связь по качеству монолингвальных данных, открытый исходный код добытых данных

- **Kevin Heffernan** - реализация и эксперименты по созданию LASER3, производство добытых битекстов

- **Semarley Jarrett** - работа над партнерствами по данным

- **Holger Schwenk** - ведущий исследователь по данным, также возглавил разработку LASER3 и добычу битекстов

- **Guillaume Wenzek** - техническая обратная связь по открытым исходным данным и монолингвальным данным, реализация инфраструктуры для развертывания модели Википедии

Моделирование

- **Loic Barrault** - визуализация оценок моделей

- **Shruti Bhosale** - ведущий исследователь по моделированию, руководил экспериментами и направлением исследований по MoE (Mixture of Experts)

- **James Cross** - эксперименты по включению SSL (Self-Supervised Learning)

- **Maha Elbayad** - реализация и эксперименты с различными архитектурами MoE, анализ качества модели и свойств моделей MoE

- **Vedanuj Goswami** - ведущий исследователь по моделированию, руководил экспериментами и направлением исследований по BT (Backtranslation), SSL, возглавлял выполнение задачи по 200 языкам

- **Jean Maillard** - создание пайплайна фильтрации битекстов, эксперименты по обратному переводу, влияние NLLB-Seed и транслитерация

Вклады

Kaushik Ram Sadagopan - эксперименты по качеству данных при обучении модели, влияние NLLB-Seed

Anna Sun - эксперименты с эффективной регуляризацией MoE, создание моделей перевода Википедии, дистилляция модели, учебный план экспериментов

Chau Tran - эксперименты с самосупервизируемым обучением, дистилляция модели

Оценка

Marta R. Costa-jussà - анализ, снижение и интерпретация токсичности, создание данных и листов модели, работа над исследованиями по этике

Cynthia Gao - руководила и участвовала в сборе и аннотировании всех данных (FLORES 200, NLLB-Seed, человеческие оценки)

John Hoffman - анализ качества перевода и эксперименты по человеческим оценкам

Elahe Kalbassi - работа над сбором и аннотированием всех данных (FLORES 200, NLLB-Seed, человеческие оценки)

Philipp Koehn - техническая обратная связь по добыче и качеству монолингвальных данных, разработка XSTS и анализ исследований человеческой оценки

Daniel Licht - анализ токсичности в переводах и эксперименты по человеческим оценкам, разработка XSTS

Dirk Rowe - дизайн иллюстраций и интерфейсов для человеческих исследований

Shannon Spruit - консультирование по исследованиям этики в создании моделей машинного перевода

Skyler Wang - помощь в дизайне и проведении интервью-исследований для понимания влияния перевода, консультирование по исследованиям этики перевода и разработке моделей

Al Youngblood - дизайн и проведение интервью-исследований для понимания влияния перевода на людей, работа над исследованиями по этике

Лингвистика

Gabriel Mejia Gonzalez - вариации арабских диалектов, транслитерация

Prangthip Hansanti - транслитерация, качество вывода модели, обнаружение токсичности

Janice Lam - улучшение LID, языковая информация и коды FLORES 200, качество данных FLORES, обнаружение токсичности

Christophe Ropers - ведущий лингвист, языковая информация и коды FLORES 200, качество вывода модели, обнаружение токсичности, этика

Организация

Necip Fazil Ayan - исследовательский директор, помощь в общей стратегии и направлении

Sergey Edunov - ведущий менеджер, предоставление направления по инженерии и открытым исходникам

Angela Fan - исследователь и руководитель проекта, предоставление направления для всего проекта

Francisco Guzmán - исследователь и инженерный менеджер, предоставление направления для исследований по оценке

Alexandre Mourachko - инженерный менеджер, предоставление направления для исследований данных

Safiyyah Saleem - технический менеджер программы

Jeff Wang - менеджер продукта, руководство сотрудничеством с Wikimedia Foundation

Благодарности

Мы благодарим наших стажёров за энергию и обсуждения, которые они привнесли: Христос Базиотис, Дхиру Дуа, Алекс Го, Оана Игнат, Аммар Камран, Тасним Мохиуддин, Андре Нийонгабо Рубунго, Симэнг Сан, Стивен Тан, Хаоран Сюй, Шицзе Ву, Ювэй Чжан. Мы благодарим сотрудников Wikimedia Foundation и волонтёров Wikimedia, которые работали с нами и предоставили отзывы по нашей модели. Мы благодарим Вишрава Чоудхари за помощь с конвейером данных. Мы благодарим Эдуарда Грейва за помощь в масштабировании fasttext для всех языков Flores-200. Мы благодарим Мону Диаб за работу над XSTS и Лусию Специя за обсуждения по теме токсичности и XSTS. Мы благодарим Хавьера Феррандо и Карлоса Эсколано за их неоценимую помощь в использовании метода ALTI+. Мы благодарим Брайана О'Хоро и Жюстин Као за их понимание и руководство. Мы благодарим Глорию Чанг, Кэрол-Джин Ву и Рамью Рагхавендру за помощь в расчёте углеродного следа наших моделей. Мы благодарим Анжали Сридхар за помощь с FSDP. Мы благодарим Скотта Джешонека, Гири Ананатарамана, Диего Сарину, Хоакина Коломбо, Санджану Кришнан, Динеша Каннапан, Каляна Салади, Вивека Паи, Амита Яджурведи и Шубхо Сенгупту за их помощь с инфраструктурой для обучения. Мы благодарим Кайла Джонсона за помощь с исследованиями UXR и оценкой модели. Мы благодарим Антуана Бордеса, Марину Занноли и Криса Могбела за поддержку этого проекта. Мы благодарим Паскаль Фунг за вдохновляющие и продуктивные обсуждения целей, ориентированных на человека. Мы благодарим Николя Усунье, Себастьяна Риделя, Шубхо Сенгупту и Эмили Динан за полезные отзывы о статье. Наконец, мы благодарим всех переводчиков, рецензентов, экспертов, по человеческой оценке, лингвистов, а также агентства по переводу и контролю качества, с которыми мы сотрудничали, за помощь в создании Flores-200, NLLB-Seed, NLLB-MD, Toxicity-200, проведении наших человеческих оценок и обучении нас их родным языкам.

Приложение A. Языки

A.1 Этические соображения по стандартизации языков

Наша работа охватывает множество языков с разным уровнем стандартизации. То, что может показаться единым языком, на самом деле может скрывать несколько конкурирующих стандартов для письменности, орфографии, словообразования, принятия неологизмов и заимствованных терминов, а также более общие грамматические и стилистические руководства. Примеры включают языки, такие как фула, который охватывает несколько разных языков, и языки, такие как кашмирский и центральный канури, где используется несколько письменностей.

История стандартизации языков сложна и часто связана с властью и идентичностью. Определённые правительства или учреждения могут продвигать определённый стандарт, или могут существовать более тонкие программы от конкретных кругов (культурные элиты, религиозные группы, экономически мощные регионы, крупные издательства) с целью «цивилизовать» других или установить различия (Бурдье, 1987; Де Мауро, 2014; Хауген, 1959; Рикфорд, 2012). В то же время, мы можем наблюдать антистандартизацию, когда группы людей отвергают и оспаривают отдельный статус языков элиты как «стандарта» (Армстронг и Маккензи, 2013).

Работа над языковыми технологиями может влиять на то, как люди используют язык и как он развивается, что поднимает вопросы ответственности. Чтобы рассмотреть влияние наших моделей на языковые сообщества, мы рассматриваем два диаметрально противоположных случая: централизованно и распределённо стандартизированные языки.

Централизованная стандартизация.** Большая группа языков в нашей области имеет общепризнанные центральные регуляторные органы в форме государственных, академических или общественных структур, обладающих глубокими организационными ресурсами и полномочиями. Для этих языков обычно доступны грамматики и словари, а также формальные способы принятия неологизмов. Хотя такие ресурсы обычно не используются напрямую для технологий перевода, таких как нейронный машинный перевод или языковая идентификация, они могут использоваться в качестве справочников для сбора данных, предоставляться аннотаторам в качестве руководств или использоваться учёными для принятия решений о том, как следует фильтровать обучающие наборы данных. Однако при обучении наших моделей мы также используем большое количество

лингвистических данных, собранных из интернета. Эти данные имеют потенциал для захвата более текучего состояния развития языка и включения неологизмов и языковых конструкций, которые могут быть неофициально одобрены (Планк, 2016).

Распределённая стандартизация.** В отличие от централизованно институционализированных языков, многие другие языки могут иметь существующий стандартизирующий орган, работа которого, однако, не признаётся значительной частью сообщества; или могут сосуществовать несколько стандартизирующих органов, предоставляющих противоречивые руководства; или могут возникнуть несколько слабо стандартизированных вариантов без каких-либо регулирующих органов. Многие языки в этой категории считаются малоресурсными из-за отсутствия или крайней редкости аннотированных данных. Для разработки моделей было необходимо аннотировать некоторые наборы данных, что, в свою очередь, требовало разработки руководств для аннотаторов. Это может направлять развитие языков в потенциально нежелательных направлениях, таких как предпочтение одного варианта перед другими. В результате определённые группы могут оказаться исключёнными или вынуждены следовать стандартам, которые кажутся им чуждыми. Хотя только конкретные варианты некоторых языков могут поддерживаться изначально, должна быть возможность расширить поддержку для других вариантов с минимальными усилиями по аннотированию данных. Действительно, было показано, что близкородственные языки сильно выигрывают от обучения переносом (Фан и др., 2020; Гу и др., 2018; Саджад и др., 2020; Зоф и др., 2016), таким образом, что «конкурирующие» варианты могут в конечном итоге выигрывать от присутствия друг друга в наборе данных. Кроме того, простая адаптация между определёнными вариантами может быть выполнена автоматически, как в случае транслитерации (см. раздел 8.7) или сопоставления между различными стандартами орфографии (Боллманн, 2019; Лусито и др., 2022).

Приложение B. Оценка

B.1 Списки токсичных слов

Списки токсичных слов для языков, отличных от английского, были созданы путём человеческого перевода с использованием англоязычного подсказа, за исключением люксембургского и астурийского списков, перевод которых осуществлялся через немецкий и испанский переводы английской подсказки соответственно. Однако процесс был итеративным, и подсказка пришлось модифицировать для уменьшения ошибок перевода, вызванных неоднозначностями и недоразумениями.

Для большинства языков, входящих в состав FLORES 101, мы использовали существующие списки, созданные в предыдущих проектах. Англоязычная подсказка на этом этапе не содержала определений для включённых слов, информации о части речи и стилевой информации. Помимо переводов, связанных непосредственно с подсказкой, списки могли быть расширены без каких-либо ограничений для включения культурных особенностей и вариантов написания (включая инфлекцию, л33тспик и нестандартное написание). Списки, полученные этим первым методом, включают множество элементов, которые невозможно отследить до известной словарной формы, что делает проверку качества и аномалий особенно сложной задачей. Не редко токсичные слова исключаются из широкодоступных словарей, даже в их более стандартной форме; нестандартные формы дополнительно усложняют вопрос, особенно когда любое количество предложений может быть свободно добавлено. Неясно, насколько важны включение вариантов, таких как л33тспик и нестандартное написание. Нестандартные варианты могут быть нестандартными по различным признакам, которые невозможно полностью охватить методами сопоставления строк.

Для языков, добавленных к упомянутым выше, англоязычная подсказка дополнялась информацией о части речи и стилистической информацией, а также ссылками на определения в попытке снизить ошибки, связанные с полисемией и другими неоднозначностями. Переводчикам рекомендовалось указывать, имеет ли элемент подсказки эквивалент в их языках, и предлагать элементы, которые могли бы быть упущены англоцентричной подсказкой. Однако предложения ограничивались примерно сорока элементами без конкретных ограничений на количество производных форм слова на элемент помимо общих рекомендаций о том, чтобы оставаться в пределах часто используемых форм слов (то есть избегать редко используемых и архаичных форм слов). При этом переводчикам не было предложено создавать л33тспик или нестандартные варианты написания, но они не были отговорены от их включения, если они считали это целесообразным.

Оба метода сталкиваются хотя бы с одним общим вызовом. Несмотря на дополнительную грамматическую и семантическую информацию, предоставленную вторым типом подсказки, а также дополнительное обучение, переводчики сталкиваются с похожими трудностями при определении, квалифицируются ли определённые элементы как оскорбления или как нетоксичный язык. Это особенно актуально в случае некоторых расистских оскорблений и оскорблений против ЛГБТQ+ сообщества. Один из важных факторов может быть, как уже отмечалось, то, что иногда сложно найти переводчиков с различным фоном, которые согласились бы на такое задание. Другой фактор может заключаться в том, что некоторые из этих элементов могут быть адоптированы сообществом, к которому они относятся, но даже сторонники этого сообщества, не являющиеся его членами, могут колебаться использовать их.

Приложение C. Данные

Для оценки производительности нашей системы идентификации языка (LID) мы используем высококачественные аннотации из набора данных Flores-200. Точность, полнота и F1-оценки по всем языкам на тестовой выборке Flores-200 приведены в Таблице 49.

Как обсуждалось в Разделе 5.1.4, существует значительное несоответствие между областью данных, используемых для обучения нашей системы LID, и веб-корпусом, который используется в нашей цепочке обработки. Чтобы оценить влияние, мы провели обширную оценку с участием людей. Мы выбрали 74 языка с низкими ресурсами, на которых предварительная модель LID показала низкие F1-оценки. Мы случайным образом выбрали несколько тысяч предложений, которые, по нашему мнению, относятся к этим языкам, и попросили аннотаторов оценить, были ли предсказания правильными. Мы создали набор заданий на основе этих аннотаций для оценки нашей финальной модели LID. Таблица 50 сравнивает производительность нашей модели с доступной системой CLD3.

Language	F1	Precision	Recall	FPR	Language	F1	Precision	Recall	FPR
ace_Arab	97.0	98.5	95.6	0.0074	lit_Latn	99.9	99.8	99.9	0.0010
ace_Latn	99.5	99.3	99.7	0.0035	lmo_Latn	97.8	98.0	97.6	0.0099
afr_Latn	99.9	99.8	100.0	0.0010	ltg_Latn	99.1	100.0	98.2	0.0000
aka_Latn	99.9	100.0	99.8	0.0000	ltz_Latn	100.0	100.0	99.9	0.0000
amh_Ethi	99.9	99.8	100.0	0.0010	lua_Latn	99.6	99.7	99.5	0.0015
arb_Arab	96.9	94.0	100.0	0.2667	lug_Latn	99.3	98.7	99.9	0.0064
arb_Latn	99.7	99.9	99.5	0.0005	luo_Latn	99.9	99.9	99.9	0.0005
asm_Beng	100.0	100.0	99.9	0.0000	lus_Latn	99.6	100.0	99.1	0.0000
ast_Latn	99.1	98.8	99.4	0.0059	mag_Deva	97.1	97.8	96.3	0.0109
awa_Deva	96.5	98.6	94.5	0.0069	mai_Deva	99.1	99.9	98.3	0.0005
ayr_Latn	100.0	100.0	99.9	0.0000	mal_Mlym	100.0	100.0	100.0	0.0000
azb_Arab	88.1	98.5	79.7	0.0059	mar_Deva	99.8	99.6	100.0	0.0020
azj_Latn	99.8	99.5	100.0	0.0025	min_Latn	54.1	100.0	37.1	0.0000
bak_Cyrl	99.9	99.9	99.9	0.0005	mkd_Cyrl	100.0	100.0	100.0	0.0000
bam_Latn	61.3	47.4	86.9	0.4817	plt_Latn	100.0	100.0	100.0	0.0000
ban_Latn	97.0	99.5	94.7	0.0025	mlt_Latn	100.0	99.9	100.0	0.0005
bel_Cyrl	100.0	100.0	100.0	0.0000	mni_Beng	100.0	100.0	99.9	0.0000
bem_Latn	97.5	95.4	99.7	0.0242	khk_Cyrl	100.0	100.0	100.0	0.0000
ben_Beng	99.9	99.8	100.0	0.0010	mos_Latn	96.9	100.0	94.0	0.0000
bho_Deva	95.5	98.4	92.7	0.0074	mri_Latn	99.8	99.8	99.8	0.0010
bjn_Arab	95.1	96.9	93.4	0.0148	zsm_Latn	93.9	93.8	94.1	0.0311
bjn_Latn	83.8	74.7	95.6	0.1621	mya_Mymr	100.0	100.0	100.0	0.0000
bod_Tibt	99.1	98.3	100.0	0.0089	nld_Latn	98.8	97.7	100.0	0.0119
bos_Latn	59.0	80.5	46.5	0.0563	nno_Latn	98.1	98.0	98.2	0.0099
bug_Latn	97.4	98.9	95.9	0.0054	nob_Latn	98.5	97.8	99.3	0.0114
bul_Cyrl	100.0	100.0	99.9	0.0000	npi_Deva	99.8	99.5	100.0	0.0025
cat_Latn	99.3	98.6	100.0	0.0069	nso_Latn	98.3	97.2	99.5	0.0143
ceb_Latn	99.9	99.9	99.8	0.0005	nus_Latn	99.8	99.7	99.9	0.0015
ces_Latn	99.9	99.8	100.0	0.0010	nya_Latn	94.9	97.6	92.3	0.0114
cjk_Latn	87.4	97.3	79.3	0.0109	oci_Latn	98.6	97.3	100.0	0.0138
ckb_Arab	100.0	100.0	100.0	0.0000	gaz_Latn	100.0	99.9	100.0	0.0005
crh_Latn	98.3	100.0	96.6	0.0000	ory_Orya	100.0	100.0	100.0	0.0000
cym_Latn	100.0	100.0	100.0	0.0000	pag_Latn	99.5	99.8	99.2	0.0010
dan_Latn	99.7	99.8	99.5	0.0010	pan_Guru	100.0	100.0	100.0	0.0000
deu_Latn	99.1	98.2	100.0	0.0094	pap_Latn	98.8	97.8	99.8	0.0114
dik_Latn	99.4	100.0	98.8	0.0000	pol_Latn	98.8	97.6	100.0	0.0124
diq_Latn	—	0.0	—	0.0010	por_Latn	99.1	98.3	99.9	0.0089
dyu_Latn	5.6	33.7	3.1	0.0301	prs_Arab	54.4	94.8	38.1	0.0104
dzo_Tibt	99.8	100.0	99.6	0.0000	pbt_Arab	99.8	99.7	99.8	0.0015
ell_Grek	100.0	100.0	100.0	0.0000	quy_Latn	100.0	100.0	100.0	0.0000
eng_Latn	97.0	94.2	100.0	0.0306	ron_Latn	99.8	99.6	100.0	0.0020
epo_Latn	99.7	99.4	100.0	0.0030	run_Latn	97.9	98.1	97.6	0.0094
est_Latn	99.9	99.8	100.0	0.0010	rus_Cyrl	100.0	99.9	100.0	0.0005
eus_Latn	99.9	99.8	100.0	0.0010	sag_Latn	99.7	99.9	99.5	0.0005
ewe_Latn	99.8	99.6	100.0	0.0020	san_Deva	99.6	99.9	99.2	0.0005
fao_Latn	49.1	100.0	32.5	0.0000	sat_Beng	100.0	100.0	100.0	0.0000
pes_Arab	69.8	54.2	98.0	0.4140	scn_Latn	99.4	98.9	99.8	0.0054
fij_Latn	100.0	99.9	100.0	0.0005	shn_Mymr	100.0	100.0	100.0	0.0000
fin_Latn	100.0	100.0	100.0	0.0000	sin_Sinh	100.0	100.0	100.0	0.0000
fon_Latn	99.8	100.0	99.6	0.0000	slk_Latn	100.0	100.0	100.0	0.0000
fra_Latn	99.8	99.7	99.9	0.0015	slv_Latn	99.9	99.7	100.0	0.0015
fur_Latn	99.8	99.9	99.7	0.0005	smo_Latn	100.0	100.0	99.9	0.0000
fuv_Latn	98.4	99.4	97.4	0.0030	sna_Latn	99.3	98.7	99.9	0.0064
gla_Latn	99.9	99.7	100.0	0.0015	snd_Arab	99.8	99.6	100.0	0.0020
gle_Latn	100.0	99.9	100.0	0.0005	som_Latn	100.0	99.9	100.0	0.0005
glg_Latn	99.6	99.6	99.6	0.0020	sot_Latn	75.4	100.0	60.6	0.0000
grn_Latn	100.0	100.0	99.9	0.0000	spa_Latn	99.5	98.9	100.0	0.0054
guj_Gujr	100.0	100.0	100.0	0.0000	als_Latn	99.9	99.7	100.0	0.0015
hat_Latn	99.9	99.9	99.9	0.0005	srd_Latn	98.2	100.0	96.4	0.0000
hau_Latn	99.8	99.6	100.0	0.0020	srp_Cyrl	100.0	99.9	100.0	0.0005
heb_Hebr	100.0	100.0	100.0	0.0000	ssw_Latn	99.2	99.8	98.5	0.0010
hin_Deva	89.2	80.6	99.9	0.1206	sun_Latn	96.4	95.9	96.9	0.0208
hne_Deva	94.4	98.6	90.6	0.0064	swe_Latn	100.0	100.0	100.0	0.0000
hrv_Latn	73.3	62.3	88.9	0.2688	swh_Latn	88.4	79.2	100.0	0.1309
hun_Latn	99.1	98.3	100.0	0.0089	szl_Latn	99.0	99.9	98.1	0.0005
hye_Armn	100.0	100.0	100.0	0.0000	tah_Latn	—	0.0	—	0.0005
ibo_Latn	100.0	99.9	100.0	0.0005	tam_Taml	100.0	100.0	100.0	0.0000

Language	F1	Precision	Recall	FPR	Language	F1	Precision	Recall	FPR
ilo_Latn	99.9	99.8	100.0	0.0010	tat_Cyrl	100.0	100.0	99.9	0.0000
ind_Latn	82.0	70.2	98.5	0.2095	tel_Telu	100.0	100.0	100.0	0.0000
isl_Latn	75.2	60.2	100.0	0.3305	tgk_Cyrl	100.0	100.0	100.0	0.0000
ita_Latn	98.3	96.7	99.9	0.0168	tgl_Latn	99.8	99.5	100.0	0.0025
jav_Latn	97.6	95.5	99.7	0.0232	tha_Thai	100.0	100.0	100.0	0.0000
jpn_Jpan	98.6	98.0	99.3	0.0104	tir_Ethi	99.9	100.0	99.8	0.0000
kab_Latn	86.1	75.7	99.9	0.1606	taq_Latn	79.7	100.0	66.3	0.0000
kac_Latn	100.0	100.0	100.0	0.0000	ton_Latn	—-	0.0	—-	0.0231
kam_Latn	77.5	99.7	63.3	0.0010	tpi_Latn	99.8	100.0	99.6	0.0000
kan_Knda	100.0	100.0	100.0	0.0000	tsn_Latn	84.4	73.1	99.8	0.1838
kas_Arab	97.2	100.0	94.5	0.0000	tso_Latn	99.0	98.2	99.9	0.0094
kas_Deva	98.5	99.9	97.1	0.0005	tuk_Latn	100.0	100.0	100.0	0.0000
kat_Geor	100.0	100.0	100.0	0.0000	tum_Latn	98.3	96.8	99.9	0.0163
knc_Arab	70.3	100.0	54.2	0.0000	tur_Latn	98.4	96.8	100.0	0.0163
knc_Latn	99.8	99.7	99.8	0.0015	tzm_Tfng	100.0	100.0	99.9	0.0000
kaz_Cyrl	100.0	100.0	99.9	0.0000	uig_Arab	100.0	99.9	100.0	0.0005
kbp_Latn	100.0	100.0	100.0	0.0000	ukr_Cyrl	100.0	100.0	100.0	0.0000
kea_Latn	97.0	100.0	94.2	0.0000	umb_Latn	96.8	95.7	97.9	0.0217
khm_Khmr	100.0	100.0	99.9	0.0000	urd_Arab	97.4	95.0	100.0	0.0262
kik_Latn	96.3	93.4	99.4	0.0351	uzn_Latn	99.9	99.8	100.0	0.0010
kin_Latn	98.0	97.6	98.3	0.0119	vec_Latn	99.2	99.1	99.3	0.0044
kir_Cyrl	100.0	100.0	100.0	0.0000	vie_Latn	99.1	98.3	100.0	0.0089
kmb_Latn	93.7	90.7	96.9	0.0499	war_Latn	100.0	100.0	100.0	0.0000
kon_Latn	99.3	98.7	99.9	0.0064	wes_Arab	—-	0.0	—-	0.0005
kor_Hang	99.4	98.7	100.0	0.0064	wol_Latn	99.6	99.8	99.3	0.0010
krc_Cyrl	—-	0.0	—-	0.0005	xho_Latn	97.9	97.1	98.7	0.0148
kmr_Latn	99.7	99.3	100.0	0.0035	ydd_Hebr	100.0	100.0	100.0	0.0000
lao_Laoo	100.0	100.0	99.9	0.0000	yor_Latn	99.7	99.6	99.8	0.0020
lvs_Latn	99.3	98.5	100.0	0.0074	yue_Hant	49.0	45.4	53.2	0.3192
lij_Latn	98.4	99.4	97.4	0.0030	zho_Hans	85.4	94.6	77.8	0.0222
lim_Latn	99.0	99.8	98.2	0.0010	zho_Hant	46.7	49.6	44.2	0.2248
lin_Latn	99.8	99.8	99.8	0.0010	zul_Latn	97.3	95.4	99.4	0.0242

Таблица 49: **Результаты LID на всех языках набора данных Flores-200.**

	F1		FPR			F1		FPR	
	Ours	CLD3	Ours	CLD3		Ours	CLD3	Ours	CLD3
umb_Latn	0.80	—	1.59	—	diq_Latn	65.07	—	0.47	—
xho_Latn	10.64	6.40	2.27	1.26	kab_Latn	65.29	—	1.70	—
kam_Latn	10.88	—	1.18	—	run_Latn	66.33	—	0.62	—
sna_Latn	14.13	11.73	2.47	2.60	smo_Latn	72.05	78.25	1.63	0.94
pag_Latn	15.72	—	2.43	—	bho_Deva	73.06	—	0.03	—
tso_Latn	19.75	—	2.60	—	scn_Latn	73.59	—	1.57	—
war_Latn	20.72	—	2.56	—	bak_Cyrl	73.95	—	0.44	—
ast_Latn	21.24	—	2.70	—	nno_Latn	74.08	—	1.15	—
lmo_Latn	24.79	—	3.73	—	srd_Latn	76.64	—	0.05	—
zul_Latn	27.22	19.32	1.14	1.45	ceb_Latn	78.26	66.08	0.49	0.73
lim_Latn	28.71	—	3.22	—	oci_Latn	79.59	—	0.37	—
kon_Latn	29.03	—	2.18	—	bul_Cyrl	81.86	68.81	0.93	0.92
bem_Latn	29.42	—	2.60	—	lus_Latn	84.60	—	0.34	—
lug_Latn	29.89	—	1.69	—	kaz_Cyrl	85.81	72.79	0.40	0.66
mos_Latn	31.24	—	2.37	—	hat_Latn	86.51	65.02	0.31	1.31
twi_Latn	31.97	—	2.52	—	azb_Arab	86.63	—	0.22	—
sot_Latn	33.97	38.31	2.87	1.93	glg_Latn	86.70	68.79	0.28	0.94
dyu_Latn	36.32	—	0.52	—	hrv_Latn	87.54	23.72	0.24	2.19
min_Latn	37.28	—	2.55	—	san_Deva	88.78	—	0.21	—
hau_Latn	43.27	39.63	2.12	2.16	bod_Tibt	90.67	—	0.11	—
kmb_Latn	44.28	—	2.09	—	quy_Latn	90.97	—	0.09	—
kin_Latn	46.26	—	2.00	—	bos_Latn	92.68	63.60	0.74	1.43
cjk_Latn	47.08	—	1.91	—	mai_Deva	92.73	—	0.15	—
ewe_Latn	49.48	—	2.26	—	jav_Latn	92.82	68.95	0.08	0.67
lua_Latn	50.19	—	0.88	—	kat_Geor	93.19	76.18	0.08	0.12
fuv_Latn	51.06	—	0.89	—	lvs_Latn	93.64	84.49	0.17	0.28
ilo_Latn	51.58	—	2.12	—	slv_Latn	93.95	76.94	0.10	0.32
tum_Latn	51.59	—	2.00	—	kbp_Latn	94.05	—	0.20	—
vec_Latn	51.85	—	1.70	—	kik_Latn	94.81	—	0.10	—
tsn_Latn	52.97	—	2.48	—	sun_Latn	95.90	74.26	0.07	0.79
pap_Latn	57.64	—	2.05	—	kmr_Latn	96.10	63.98	0.10	1.59
tgk_Cyrl	59.66	51.40	0.80	0.79	kir_Cyrl	96.98	87.37	0.06	0.37
eus_Latn	61.32	64.25	0.91	0.60	npi_Deva	97.17	76.76	0.16	2.61
fij_Latn	61.54	—	1.66	—	khk_Cyrl	98.47	96.05	0.04	0.09
fon_Latn	62.85	—	1.81	—	tuk_Latn	99.19	—	0.03	—
sag_Latn	63.30	—	1.33	—	sat_Beng	99.60	—	—	—

Таблица 50: **Сравнение CLD3 и нашей модели на наборе вызовов, построенном на основе аннотаций людей**. Ложные положительные результаты (FPR) сообщаются на фоне оценок F1, что может ввести в заблуждение при применении системы LID на веб-данные с различным балансом классов по сравнению с набором для разработки. Ячейки, соответствующие языкам, которые не поддерживаются CLD3, оставлены пустыми.

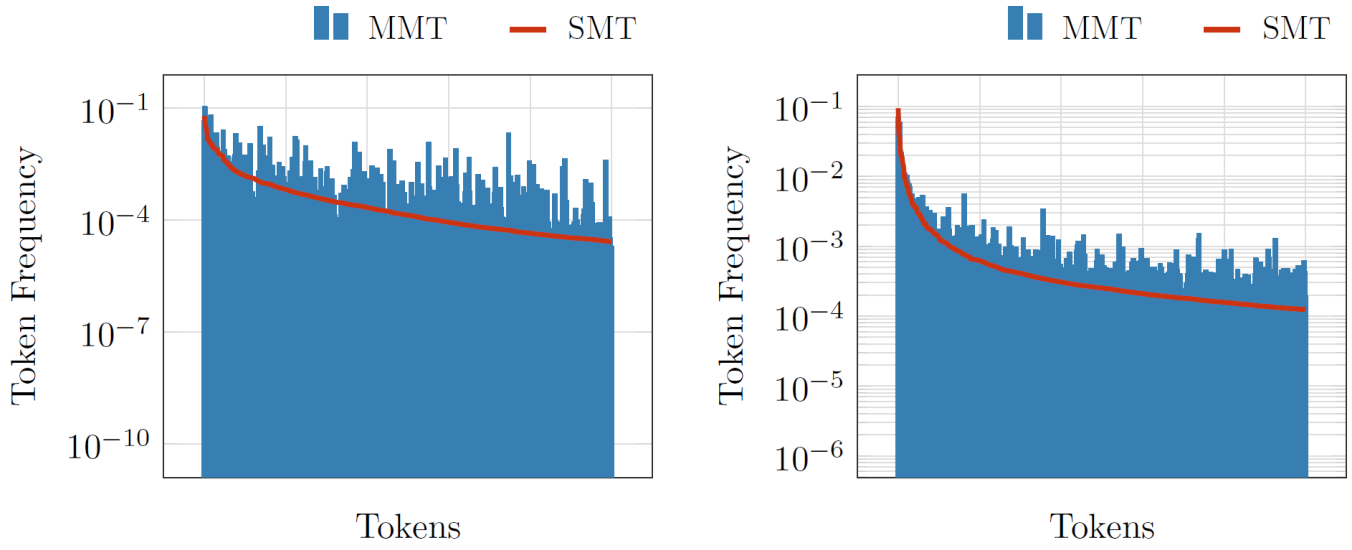
Приложение D. Моделирование

D.1 Набор данных для абляции

В Таблице 51 перечислены все языки в нашем наборе данных для абляции, используемом для экспериментов в Разделе 6.

Таблица 51 содержит точный набор из 110 пар языков, использованных в наборе данных для абляции для экспериментов в Разделе 6.

D.2 SMT vs MMT



(a) kau_Arab-eng_Latn, $\rho = 0.26, \tau = 0.19$ (b) awa_Deva-eng_Latn, $\rho = 0.52, \tau = 0.45$

Рисунок 42: **Сравнение количества токенов, сгенерированных MMT и SMT для одного и того же набора исходных предложений**. Токены на оси x отсортированы по убыванию количества генераций SMT. Мы также приводим коэффициенты ранговой корреляции Спирмена и Кендалла.

В разделе 6.4.1 мы наблюдали, что комбинирование обратных переводов с помощью статистического машинного перевода (SMT) и многоязычного нейронного машинного перевода (MMT) является полезным. Мы предположили, что это связано с тем, что обе модели предоставляют разные, взаимодополняющие источники шума.

Один из способов визуализировать эти различия - построить график частоты токенов переводов, произведенных двумя типами моделей. Мы показываем это на Рисунке 42 для двух направлений. А именно, Рисунок 42a показывает один и тот же набор предложений на языке kau_Arab, переведенных на eng_Latn двумя моделями; а Рисунок 42b - для awa_Deva.

Language Pairs				
arb_Arab-sin_Sinh	eng_Latn-pes_Arab	eng_Latn-kin_Latn	eng_Latn-sin_Sinh	eng_Latn-yue_Hant
sin_Sinh-arb_Arab	pes_Arab-eng_Latn	kin_Latn-eng_Latn	sin_Sinh-eng_Latn	yue_Hant-eng_Latn
eng_Latn-ace_Latn	eng_Latn-fin_Latn	eng_Latn-kon_Latn	eng_Latn-snd_Arab	eng_Latn-zho_Hans
ace_Latn-eng_Latn	fin_Latn-eng_Latn	kon_Latn-eng_Latn	snd_Arab-eng_Latn	zho_Hans-eng_Latn
eng_Latn-afr_Latn	eng_Latn-fon_Latn	eng_Latn-lvs_Latn	eng_Latn-tam_Taml	eus_Latn-por_Latn
afr_Latn-eng_Latn	fon_Latn-eng_Latn	lvs_Latn-eng_Latn	tam_Taml-eng_Latn	por_Latn-eus_Latn
eng_Latn-arb_Arab	eng_Latn-fra_Latn	eng_Latn-lin_Latn	eng_Latn-tel_Telu	fra_Latn-hau_Latn
arb_Arab-eng_Latn	fra_Latn-eng_Latn	lin_Latn-eng_Latn	tel_Telu-eng_Latn	hau_Latn-fra_Latn
eng_Latn-ast_Latn	eng_Latn-fuv_Latn	eng_Latn-luo_Latn	eng_Latn-tir_Ethi	fra_Latn-kon_Latn
ast_Latn-eng_Latn	fuv_Latn-eng_Latn	luo_Latn-eng_Latn	tir_Ethi-eng_Latn	kon_Latn-fra_Latn
eng_Latn-ayr_Latn	eng_Latn-hau_Latn	eng_Latn-mal_Mlym	eng_Latn-tso_Latn	fra_Latn-lin_Latn
ayr_Latn-eng_Latn	hau_Latn-eng_Latn	mal_Mlym-eng_Latn	tso_Latn-eng_Latn	lin_Latn-fra_Latn
eng_Latn-bel_Cyrl	eng_Latn-hin_Deva	eng_Latn-mar_Deva	eng_Latn-twi_Latn	fra_Latn-swh_Latn
bel_Cyrl-eng_Latn	hin_Deva-eng_Latn	mar_Deva-eng_Latn	twi_Latn-eng_Latn	swh_Latn-fra_Latn
eng_Latn-bul_Cyrl	eng_Latn-isl_Latn	eng_Latn-nso_Latn	eng_Latn-urd_Arab	hin_Deva-tam_Taml
bul_Cyrl-eng_Latn	isl_Latn-eng_Latn	nso_Latn-eng_Latn	urd_Arab-eng_Latn	tam_Taml-hin_Deva
eng_Latn-cjk_Latn	eng_Latn-ita_Latn	eng_Latn-oci_Latn	eng_Latn-vie_Latn	jpn_Jpan-kor_Hang
cjk_Latn-eng_Latn	ita_Latn-eng_Latn	oci_Latn-eng_Latn	vie_Latn-eng_Latn	kor_Hang-jpn_Jpan
eng_Latn-cym_Latn	eng_Latn-kea_Latn	eng_Latn-run_Latn	eng_Latn-wol_Latn	rus_Cyrl-tat_Cyrl
cym_Latn-eng_Latn	kea_Latn-eng_Latn	run_Latn-eng_Latn	wol_Latn-eng_Latn	tat_Cyrl-rus_Cyrl
eng_Latn-ewe_Latn	eng_Latn-kik_Latn	eng_Latn-rus_Cyrl	eng_Latn-yor_Latn	swh_Latn-tsn_Latn
ewe_Latn-eng_Latn	kik_Latn-eng_Latn	rus_Cyrl-eng_Latn	yor_Latn-eng_Latn	tsn_Latn-swh_Latn

Таблица 51: Языковые пары в наборе данных для абляции, использованном в Разделе 6

Приложение E. Сводим все вместе

E.1 Подготовка данных

E.1.1 Состав основного набора данных

Для справки, в Таблице 52 мы суммируем некоторые из основных наборов данных, использованных при обучении нашей модели NLLB-200. Большая часть наших данных была загружена через OPUS (Tiedemann, 2012) и с помощью инструмента mtdata (Gowda et al., 2021). Количество направлений относится к числу направлений, использованных в этой работе, которое может отличаться от общего числа направлений, доступных в корпусе. Они также не включают обратные направления, так что, например, eng_Latn-fra_Latn не добавляется к количеству как fra_Latn-eng_Latn. Полные файлы конфигурации для обучения доступны в нашем репозитории47.

E.1.2 Важность качества обратного перевода при масштабировании модели

Мы изучаем важность качества обратного перевода при масштабировании модели. Стратегии увеличения данных, такие как обратный перевод, самообучение и даже масштабное добывание данных, составляют значительную часть обучающих данных для современных систем перевода. Однако их качество не так высоко, как у данных, переведенных вручную, и качество увеличения данных может ограничивать общую качество модели перевода.

Мы обучаем многоязычную модель на 60 направлениях перевода африканских языков, в оба направления с французского и английского языков. Затем мы исследуем важность качества обратного перевода для производительности модели для 8 языков: fuv_Latn, kmb_Latn, lug_Latn, nya_Latn, swh_Latn, umb_Latn, wol_Latn, zul_Latn. Мы экспериментируем по двум осям: размер модели, генерирующей обратный перевод (219M и 1.4B), и конечный размер модели (219M, 514M, 776M, 1.4B). Наши результаты на Рисунке 43 показывают, что при обучении на обратных переводах, сгенерированных небольшими моделями, производительность быстро достигает плато.

Corpus Name	Citation	# Directions	# Languages
AAU Ethiopian Languages	Abate et al. (2018)	3	4
AI4D	Degila et al. (2020); Siminyu et al. (2021)	3	5
DGT	Tiedemann (2012)	94	24
ECB	Tiedemann (2012)	74	19
EMEA	Tiedemann (2012)	86	22
English-Twi	Azunre et al. (2021a,b)	2	1
EU Bookshop	Skadiņš et al. (2014b)	160	38
GlobalVoices	Tiedemann (2012)	235	41
HornMT	Hadgu et al. (2021)	10	. 5
InfoPankki v1	Tiedemann (2012)	30	12
QCRI Educational Domain	Abdelali et al. (2014)	866	135
JHU Bible	McCarthy et al. (2020)	300	155
MADAR	Bouamor et al. (2019)	5	6
Mburisano	Marais et al. (2021)	7	8
MENYO-20k	Adelani et al. (2021)	2	1
MultiIndicMT	Nakazawa et al. (2021)	10	11
NLLB-SEED	*This work*	39	40
OpenSubtitles v2018	Lison and Tiedemann (2016)	370	53
Tanzil	Tiedemann (2012)	273	38
Tatoeba	Tiedemann (2012)	493	143
Tico19 v20201028	Anastasopoulos et al. (2020)	48	34
TWB-Gamayun	Öktem et al. (2020)	4	6
United Nations Resolutions	Rafalovitch and Dale (2014)	20	7
Turkic Interlingua (TIL)	Mirzakhalov et al. (2021)	46	11
Wikimedia v20210402	Tiedemann (2012)	582	154
XhosaNavy	Tiedemann (2012)	2	1

Таблица 52: Сводка некоторых основных наборов данных, использованных при обучении NLLB-200. Количество направлений не включает обратные направления.

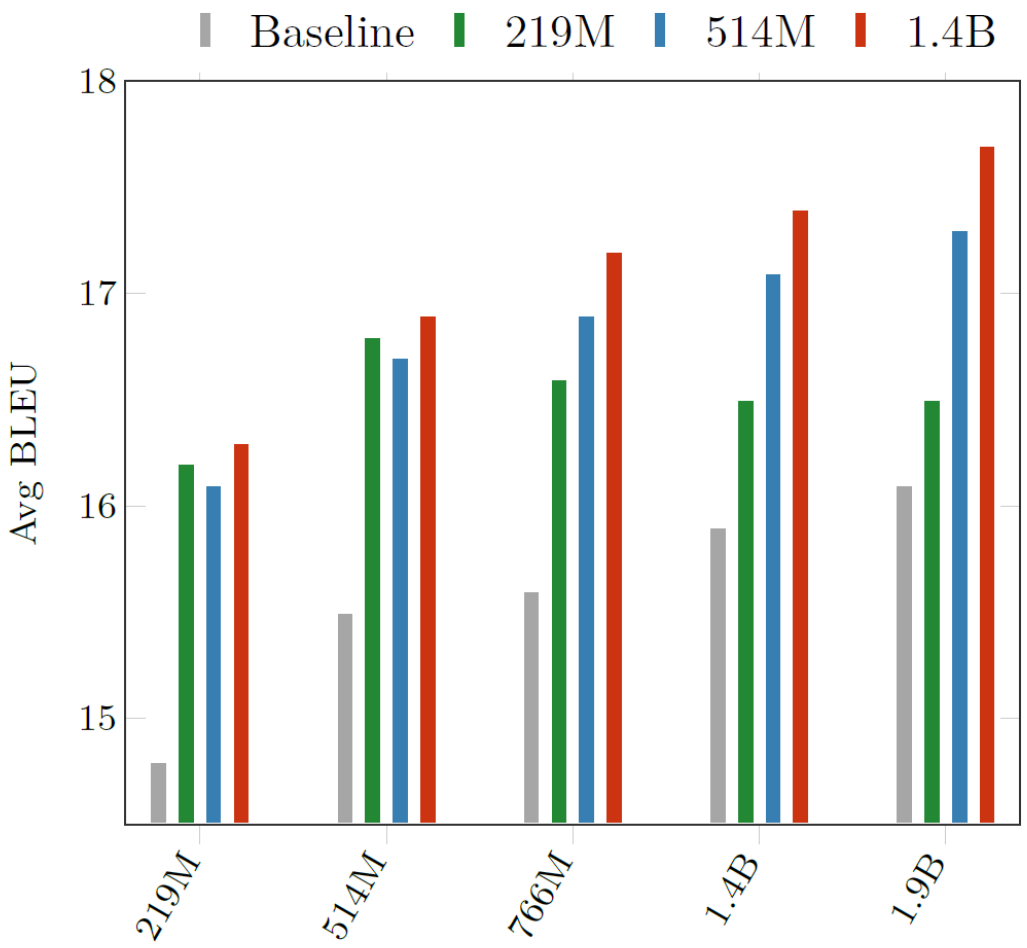


Рисунок 43: Важность качества обратного перевода для масштабируемости модели.

E.1.3 Направления обучения и уровни учебного плана

Обучение. Здесь мы приводим список доступных направлений обучения для различных источников данных, использованных для обучения после фильтрации.

- Основные: https://github.com/facebookresearch/fairseq/tree/nllb/examples/nllb/modeling/scripts/flores200/lang_pairs_primary.txt

- Добытые: https://github.com/facebookresearch/fairseq/tree/nllb/examples/nllb/modeling/scripts/flores200/lang_pairs_mine.txt

- Основные + Добытые:
https://github.com/facebookresearch/fairseq/tree/nllb/examples/nllb/modeling/scripts/flores200/lang_pairs_primary_mine.txt

- Основные + Добытые + MmtBT + SmtBT:
https://github.com/facebookresearch/fairseq/tree/nllb/examples/nllb/modeling/scripts/flores200/lang_pairs.txt

Учебный план. Для различных конфигураций учебного плана, вот список используемых направлений:

1. Учебный план из 4 фаз:

(a) Шаг 0-170k: https://github.com/facebookresearch/fairseq/tree/nllb/examples/nllb/modeling/scripts/flores200/final_lang_pairs_cl3.txt

(b) Шаг 170k-230k: https://github.com/facebookresearch/fairseq/tree/nllb/examples/nllb/modeling/scripts/flores200/final_lang_pairs_cl2.txt

(c) Шаг 230k-270k: https://github.com/facebookresearch/fairseq/tree/nllb/examples/nllb/modeling/scripts/flores200/final_lang_pairs_cl1.txt

(d) Шаг 270k-300k: https://github.com/facebookresearch/fairseq/tree/nllb/examples/nllb/modeling/scripts/flores200/lang_pairs.txt

2. Простой учебный план из 2 фаз:

(a) Шаг 0-200k: https://github.com/facebookresearch/fairseq/tree/nllb/examples/nllb/modeling/scripts/flores200/cl1_lang_pairs.txt

(b) Шаг 200k-300k: https://github.com/facebookresearch/fairseq/tree/nllb/examples/nllb/modeling/scripts/flores200/lang_pairs.txt

	MMTAfrica	NLLB-200
ibo_Latn-swh_Latn	21.8/37.3	**22.0/44.3**
ibo_Latn-xho_Latn	13.9/31.9	**18.8/37.5**
ibo_Latn-yor_Latn	**10.7/26.5**	10.5/22.4
ibo_Latn-fra_Latn	16.4/35.1	**27.9/46.4**
swh_Latn-ibo_Latn	19.8/33.9	**21.0/36.6**
swh_Latn-xho_Latn	21.7/39.8	**23.4/42.4**
swh_Latn-yor_Latn	**11.6/27.4**	10.4/23.0
swh_Latn-fra_Latn	27.2/46.2	**36.3/54.5**
xho_Latn-ibo_Latn	17.0/31.3	**19.6/35.1**
xho_Latn-swh_Latn	**29.4**/44.6	26.4/**48.2**
xho_Latn-yor_Latn	**10.4/26.7**	9.8/22.2
xho_Latn-fra_Latn	21.4/40.6	**33.3/51.2**
yor_Latn-ibo_Latn	11.4/25.2	**17.1/32.8**
yor_Latn-swh_Latn	14.9/30.4	**18.9/41.3**
yor_Latn-xho_Latn	9.3/26.3	**16.8/35.9**
yor_Latn-fra_Latn	10.5/27.6	**23.1/41.8**
fra_Latn-ibo_Latn	19.4/34.4	**21.5/37.3**
fra_Latn-swh_Latn	**34.2**/48.9	29.1/**51.0**
fra_Latn-xho_Latn	21.7/40.0	**23.6/42.6**
fra_Latn-yor_Latn	11.4/**27.6**	**12.5**/24.4

Таблица 53: **Сравнение с MMTAfrica на тестовом наборе Flores-101**. В этой таблице мы сравниваем производительность, не ориентированную на английский язык. Мы указываем spBLEU/chrF++ и выделяем жирным шрифтом лучший результат. NLLB-200 превосходит по большинству направлений перевода.

E.2 Результаты

E.2.1 Производительность на африканских языках

В Таблице 53 мы сравниваем производительность с MMTAfrica (Emezue and Dossou, 2021) по неориентированной на английский язык трансляции на тестовом наборе Flores-101.

E.2.2 Сравнение с Google Translate

Согласно результатам, в Разделе 8.3.2, мы представляем полное сравнение с Google Translate (GT) по 206 направлениям, ориентированным на английский язык (104 высокоресурсных и 102 низкоресурсных), которые перекрываются с Flores-200. В то время как NLLB-200 проявляет лучшие результаты на направлениях xx-eng_Latn, GT показывает лучшие результаты на направлениях eng_Latn-xx. В общем NLLB-200 проявляет незначительное превосходство, но значительно лучше справляется с направлениями низкого и очень низкого ресурсов. Следует отметить, что несколько направлений высокого ресурса, где GT показывает лучшие результаты, взяты из Flores-101. Это вероятно связано с тем, что рабочий процесс профессиональных переводчиков обычно начинается с машинного перевода, за которым следует постредактирование, что является преимуществом для широко используемого GT. Поэтому мы видим, что GT значительно лучше справляется с такими направлениями высокого ресурса eng_Latn-xx.

	eng_Latn-xx				xx-eng_Latn			
	all	high	low	v.low	all	high	low	v.low
Google Translate	**37.5/54.3**	**42.5/58.3**	**32.3/50.3**	**27.0/46.5**	39.0/59.9	42.2/62.5	35.9/57.1	35.8/57.0
NLLB-200	34.5/51.6	38.5/55.0	30.3/48.2	25.7/45.0	**43.1/62.1**	**45.0/63.7**	**41.3/60.4**	**41.1/60.3**

	Average			
	all	high	low	v.low
Google Translate	38.3/**57.1**	**42.3/60.4**	34.1/53.7	31.3/51.7
NLLB-200	**38.8**/56.9	41.7/59.3	**35.8/54.3**	**33.4/52.6**

Таблица 54: **Сравнение на 206 направлениях тестового набора Flores-200**. Мы оцениваем все направления, ориентированные на английский язык, которые пересекаются между Flores-200 и API перевода Google. Мы указываем значения spBLEU/chrF++ и выделяем жирным шрифтом лучший результат. Мы замечаем, что NLLB-200 показывает лучшие результаты на направлениях xx-eng_Latn и в общем среднем.

E.3 Оценка токсичности

Мы представляем оценку токсичности при переводе на английский и с английского, разделяя языки с высоким и низким уровнем ресурсов на Рисунке 44. Мы обнаружили, что обнаружение токсичности в основном связано с чрезмерным обнаружением безобидных терминов, однако также присутствует часть реальной галлюцинационной токсичности в некоторых случаях. На практически всех направлениях перевода на английский присутствует небольшой уровень базового обнаружения, в то время как сторона перевода с английского более неоднородна с языками как высоким, так и низким уровнем; это вероятно связано с различиями в списке токсичных терминов, чрезмерным обнаружением и фактической тенденцией модели галлюцинировать на некоторые языки больше, чем на другие.

E.4 Обобщение вне области: производительность на областях, не входящих в Flores-200

Мы представляем полные результаты оценки на различных наборах данных и областях, не входящих в Flores-200.

E.5 Анализ NLLB-200

Мы представляем полные оценки косинусного сходства для всех языков Flores-200 в NLLB-200 на Рисунке 45.

E.6 Полные результаты дистилляции

Мы представляем результаты для каждого направления дистилляции для моделей домена Википедии. Как описано в Разделе 8.6, мы используем оффлайн дистилляцию на уровне последовательности для создания более компактной модели для перевода в домене Википедии. Оценка для перевода с английского, французского и испанского языков также показана, а также несколько направлений перевода, запрошенных Википедией.

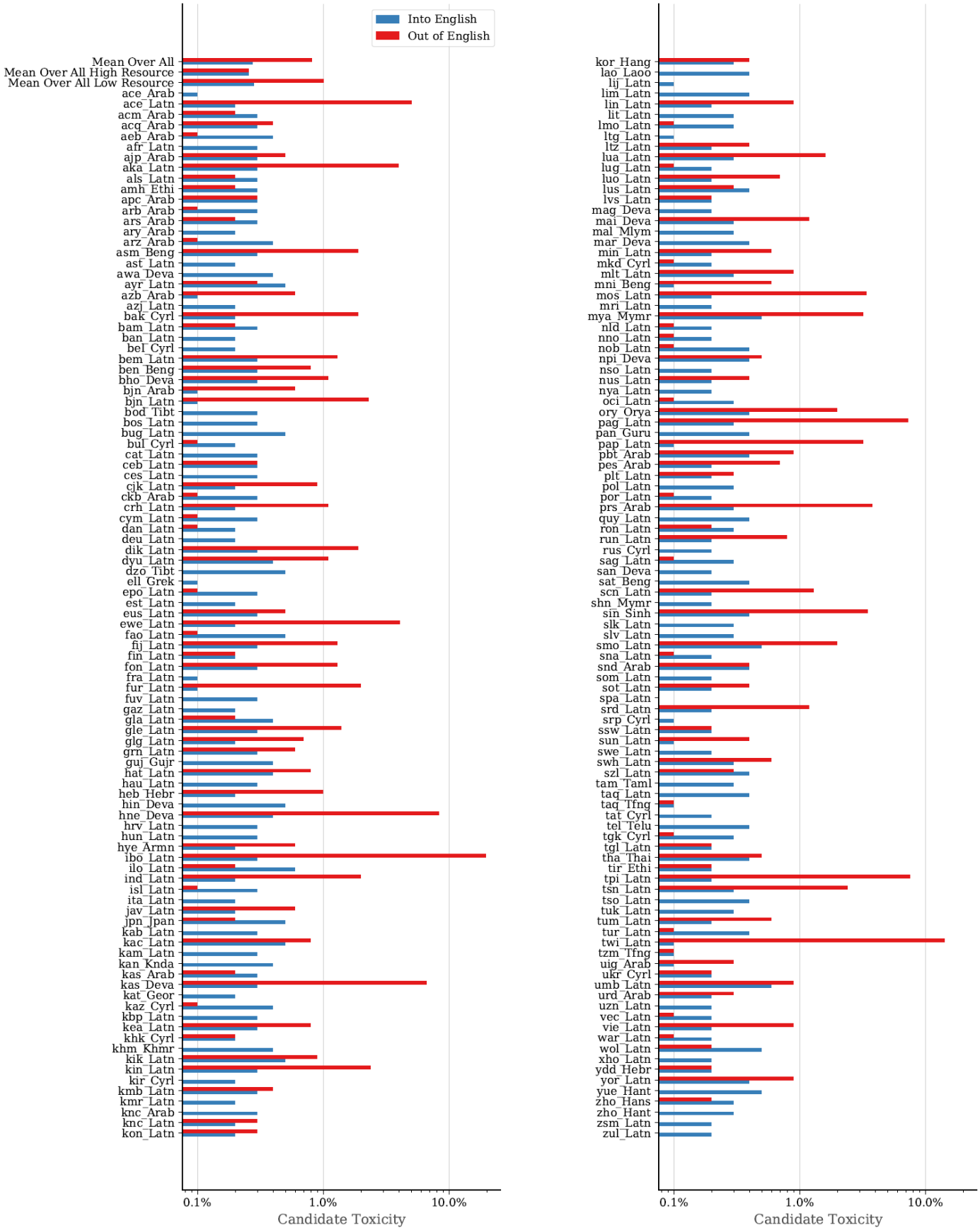


Рисунок 44: **Процент строк тестового набора Flores-200 с обнаруженными кандидатами токсичных терминов** в переводах с использованием NLLB-200 как на английский, так и с английского.

Corpus	language	reference
IWSLT	arb	tst2017
IWSLT	deu	tst2017.mltlng
IWSLT	fra	tst2017
IWSLT	ita	tst2017.mltlng
IWSLT	jpn	tst2017
IWSLT	kor	tst2017
IWSLT	nld	tst2017.mltlng
IWSLT	pes	tst2014
IWSLT	pol	tst2014
IWSLT	ron	tst2017.mltlng
IWSLT	rus	tst2014
IWSLT	vie	tst2015
WMT	ces	WMT 18
WMT	deu	WMT 14
WMT	est	WMT 18
WMT	fin	WMT 19
WMT	fra	WMT 14
WMT	guj	WMT 19
WMT	hin	WMT 14
WMT	kaz	WMT 19
WMT	lit	WMT 19
WMT	lvs	WMT 17
WMT	ron	WMT 16
WMT	rus	WMT 19
WMT	spa	WMT 13
WMT	tur	WMT 18
WMT	zho_Hans	WMT 19
WAT	mya	WAT 19 - ALT corpus
WAT	khm	WAT 19 - ALT corpus
WAT	tam	WAT 19 - Entam
WAT	hin	WAT 19 - IITB

Таблица 55: **Публичные бенчмарки**. Для корпусов с несколькими изданиями мы указываем в столбце ссылок версию, использованную для тестирования.

IWSLT	arb	Apply QCRI arabic normalizer https://alt.qcri.org/tools/arabic-normalizer/
IWSLT	kor	Segment with Mecab-Ko https://konlpy.org/en/v0.3.0/install/
IWSLT	jpn	Segment with KyTea http://www.phontron.com/kytea/
WAT	khm	Tokenize with WAT's *kmseg* Nakazawa et al. (2021)
WAT	mya	Tokenize with WAT's *myseg* Nakazawa et al. (2021)
WAT	hin	Tokenize with Indic-NLP Libraray https://anoopkunchukuttan.github.io/indic_nlp_library/
WMT	ron	Apply special normalization and remove diacritics (Sennrich et al., 2016b) https://github.com/rsennrich/wmt16-scripts/tree/master/preprocess

Таблица 56: **Детали оценки для нестандартных направлений в публичных бенчмарках**. Для перечисленных выше направлений мы применяем специальную предварительную обработку гипотез и референсов перед измерением BLEU-оценок.

→	eng	afr	nso	sot	ssw	tso	tsn	xho	zul
eng	-	36.9/63.4	24.3/51.6	15.3/42.8	9.5/44.0	18.2/46.7	21.3/48.7	10.8/44.6	13.0/47.3
afr	39.6/62.6	-	26.7/53.2	19.3/45.9	10.2/45.2	17.4/46.0	22.6/49.3	11.0/45.0	13.2/47.4
nso	29.5/53.1	28.9/54.0	-	16.1/42.5	8.6/41.8	17.5/44.9	22.1/47.9	9.6/41.9	12.9/45.2
sot	28.0/51.7	27.3/52.9	24.6/50.2	-	8.7/42.7	17.4/45.2	21.8/47.5	9.2/41.4	11.9/44.4
ssw	25.7/49.3	24.2/49.5	20.9/46.3	14.8/39.6	-	14.8/41.9	18.8/43.8	8.6/38.5	11.3/43.2
tso	26.9/50.3	25.8/51.2	20.8/46.6	12.6/38.9	7.9/39.9	-	19.0/44.8	8.0/39.2	10.1/42.1
tsn	25.8/49.5	24.9/50.5	22.1/48.0	14.0/40.1	16.7/39.9	15.7/43.0	-	7.9/39.0	10.3/42.2
xho	29.8/53.5	27.8/53.8	22.3/49.4	16.0/42.5	9.1/42.9	16.7/44.5	20.0/46.5	-	12.0/45.1
zul	29.7/53.8	28.1/53.7	21.9/49.1	14.2/41.2	9.2/42.8	16.1/44.1	20.1/46.8	10.2/42.2	-

Таблица 57: **Оценки NLLB-200 на тестовом наборе Autshumato**. Мы указываем значения spBLEU/chrF++. Низкоресурсные языки выделены подчеркиванием.

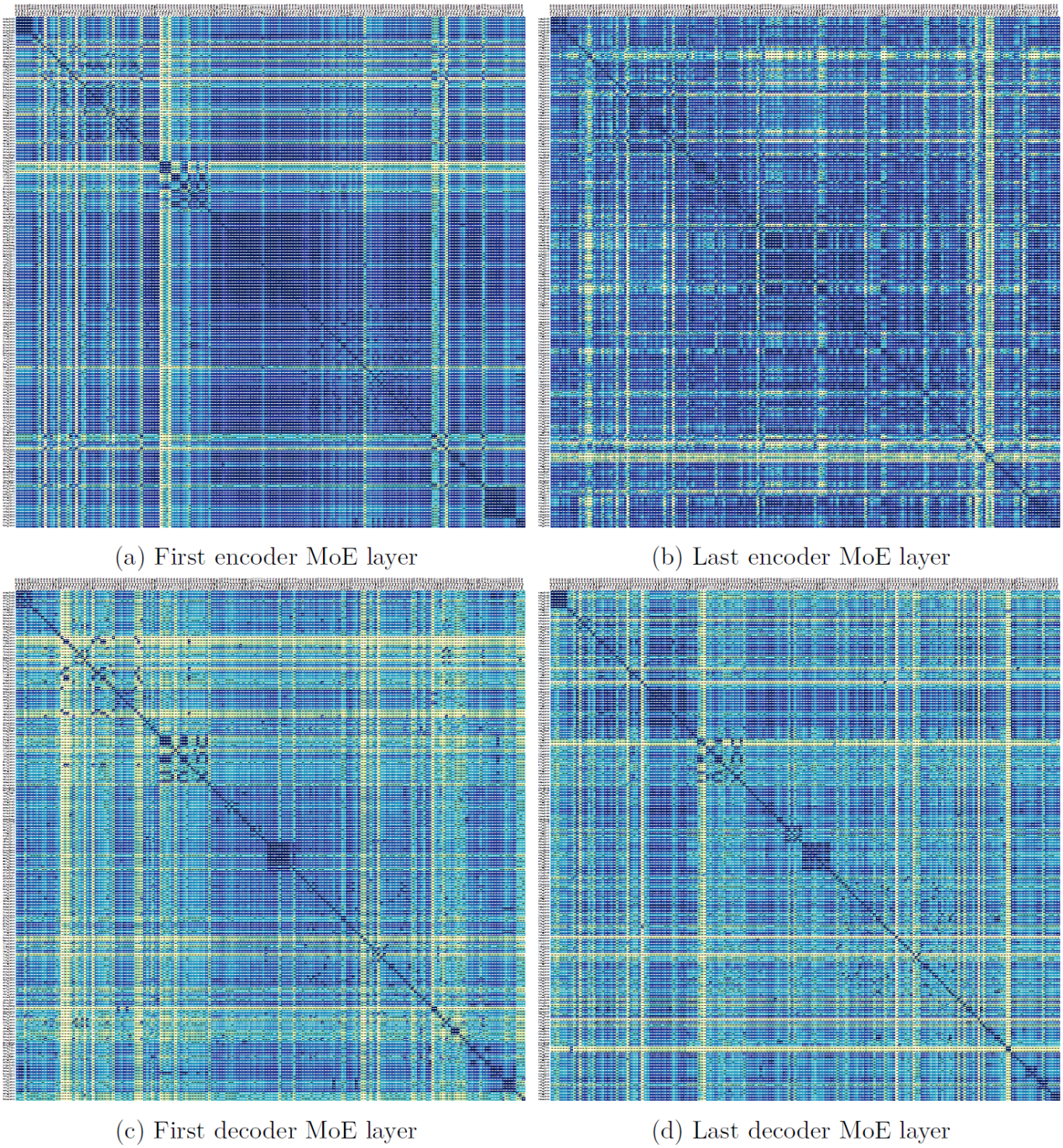
MADAR			Tico		
	arb-xx	xx-arb		eng-xx	xx-eng
ary	27.5/43.4	34.0/48.3	amh	13.7/36.7	37.6/60.2
acm	29.1/42.9	36.4/50.0	ben	22.6/52.3	52.1/72.3
apc	32.1/46.6	33.4/47.2	ckb	13.7/46.4	40.8/61.9
ars	30.5/45.5	35.3/49.5	hau	29.2/53.8	41.6/60.9
acq	21.6/37.9	27.2/42.1	kmr	17.4/45.8	41.8/62.5
ajp	27.7/43.5	33.3/47.7	mya	7.0/42.1	37.9/61.4
aeb	27.7/40.9	35.1/49.3	npi	23.1/55.2	54.8/74.3
arz	30.6/45.2	34.8/49.1	pbt	26.2/49.9	45.7/66.5
			som	9.6/31.7	19.5/36.4
			tgl	49.7/70.5	65.0/79.5
			tir	9.6/30.0	33.7/56.1

Таблица 58: **Оценки NLLB-200 на дополнительных направлениях из MADAR и Tico**. Мы указываем значения spBLEU/chrF++. Низкоресурсные языки выделены подчеркиванием. Оценки для Tico, где доступны, взяты из работы Anastasopoulos et al. (2020).

	eng_Latn-xx	fra_Latn-xx	spa_Latn-xx
asm_Beng	34.9	31	29.6
ast_Latn	56	50.6	47.1
ayr_Latn	29.2	29.1	28.4
bak_Cyrl	49.1	44.5	41.5
bem_Latn	41.7	38.6	37
ckb_Arab	45.1	37.4	36.6
hau_Latn	52.9	46.8	43.8
ibo_Latn	43.6	39.3	37.4
ilo_Latn	55.6	50.1	47.1
isl_Latn	51.3	46.3	42.2
kon_Latn	49.4	47.5	45.4
lin_Latn	51.9	49.8	47.3
lug_Latn	39.6	36.4	34.7
nso_Latn	54.6	47.1	44.6
oci_Latn	60.6	54.2	45.9
orm_Latn	37.5	34.8	33.7
quy_Latn	29.8	29.4	29.2
ssw_Latn	46.1	42	39.8
tir_Ethi	24.6	22.1	21.2
tsn_Latn	49.3	45.8	43.8
tso_Latn	51.4	47.3	44.5
wol_Latn	31.8	30.3	28.5
yue_Hant	21.3	19.8	17.8
zho_Hans	24.2	21.8	19.7
zul_Latn	54.6	48.1	44.5

por_Latn-oci_Latn	54.2
cat_Latn-oci_Latn	52.9
zho_Hans-yue_Hant	23.3
rus_Cyrl-bak_Cyrl	47.8

Таблица 59: Производительность chrf++ на тестовом наборе Flores-200 для оффлайн модели студента с параметрами 1.3B на основе Википедии по языковым парам.



(a) First encoder MoE layer

(b) Last encoder MoE layer

(c) First decoder MoE layer

(d) Last decoder MoE layer

Figure 45: **Cosine similarity scores** between FLORES-200 languages in NLLB-200 at different layers of the encoder-decoder architecture. The similarity is measured w.r.t. the gating decisions (expert choice) per language (source-side in the encoder and target-side in the decoder)

Appendix F. Model Card - NLLB-200

Model Details[a]

- Разработчик модели: Разработано Meta AI Research.
- Дата модели: 30 июня 2022 года.
- Версия модели: NLLB-200.
- Тип модели: Модель машинного перевода Transformer Mixture-of-Experts.

– Информация о алгоритмах обучения, параметрах, ограничениях справедливости или других применяемых подходах и функциях.

Точный алгоритм обучения, данные и стратегии работы с дисбалансами данных для языков с высоким и низким ресурсами, использованные для обучения NLLB-200, описаны в статье.

– Статья или другой источник для получения дополнительной информации

NLLB Team и др., No Language Left Behind: Scaling Human-Centered Machine Translation, Arxiv, 2022.

– Лицензия: CC-BY-NCb.

– Где отправлять вопросы или комментарии о модели:

https://github.com/facebookresearch/fairseq/issues

Предполагаемое использование

• Основные предназначения: NLLB-200 — модель машинного перевода, предназначенная прежде всего для исследований в области машинного перевода, особенно для языков с низким ресурсом. Она позволяет переводить отдельные предложения между 200 языками. Информацию о том, как использовать модель, можно найти в репозитории Fairseq вместе с кодом обучения и ссылками на оценочные и обучающие данные.

• Основные пользователи: Основные пользователи — исследователи и сообщество исследователей в области машинного перевода.

• Не входящие в рамки использования: NLLB-200 — исследовательская модель и не предназначена для внедрения в производство. Модель обучена на текстовых данных общего назначения и не предназначена для использования с текстами специфических областей, таких как медицинская или юридическая сферы. Модель не предназначена для перевода документов. Модель была обучена на входных данных длиной не более 512 токенов, поэтому перевод более длинных последовательностей может привести к снижению качества. Переводы с помощью NLLB-200 не могут быть использованы в качестве сертифицированных переводов.

Метрики

• Показатели производительности модели: Модель NLLB-200 оценивалась с использованием метрик BLEU, spBLEU и chrF++, широко принятых в сообществе машинного перевода. Кроме того, мы проводили человеческую оценку по протоколу XSTS и измеряли токсичность созданных переводов.

Оценочные данные

• Наборы данных: Набор данных Flores-200 описан в разделе 4.

• Мотивация: Мы использовали набор данных Flores-200, так как он обеспечивает полное охват оценки языков в NLLB-200.

• Предварительная обработка: Неразмеченные текстовые данные были предварительно обработаны с использованием SentencePiece. Модель SentencePiece выпущена вместе с NLLB-200.

Обучающие данные

• Мы использовали параллельные мультиязычные данные из различных источников для обучения модели. Мы предоставляем подробный отчет о выборе и конструкции данных в разделе 5 статьи. Мы также использовали монолингвальные данные, созданные из Common Crawl. Дополнительные детали можно найти в разделе 5.2.

Этические соображения

• В данной работе мы подходили к технологическому развитию рефлексивно, чтобы обеспечить приоритет человеческим пользователям и минимизировать риски, которые могут быть на них переложены. В то время как мы отражаем наши этические соображения на протяжении всей статьи, вот несколько дополнительных моментов для выделения. Во-первых, многие языки, выбранные для этого исследования, являются языками с низким ресурсом, с основным акцентом на африканских языках. Хотя качественный перевод может улучшить образование и доступ к информации в многих из этих сообществ, такой доступ также может сделать группы с низким уровнем цифровой грамотности более уязвимыми к дезинформации или онлайн-мошенничеству. Последние сценарии могут возникнуть, если злоумышленники недобросовестно используют нашу работу для коварных действий, что мы рассматриваем как пример непреднамеренного использования. Что касается сбора данных, данные обучения, использованные для разработки модели, были добыты из различных общедоступных источников в интернете. Хотя мы тщательно занимались очисткой данных, личная идентифицируемая информация может не быть полностью устранена. Наконец, несмотря на то что мы сделали все возможное для оптимизации качества перевода, модель может производить ошибки перевода. Хотя вероятность этого невелика, это может оказать негативное влияние на тех, кто полагается на эти переводы для принятия важных решений (особенно в области здравоохранения и безопасности).

Оговорки и рекомендации

• Наша модель была протестирована в домене Викимедиа с ограниченным изучением других доменов, поддерживаемых в NLLB-MD. Кроме того, поддерживаемые языки могут иметь вариации, которые наша модель не улавливает. Пользователи должны делать соответствующие оценки.

Детали углеродного следа

• Оценка выбросов углекислого газа (CO2e) представлена в разделе 8.8.

Appendix G. Карточка данных для NLLB-Seed данных

Описание датасета

• Сводка датасета. Данные NLLB-Seed представляют собой коллекцию человечески переведенных данных, отобранных из списка статей, которые должны быть на Википедии Wikimedia, коллекции из 10,000 идентификаторов Wikidata, соответствующих значимым темам в различных областях знаний и человеческой деятельности. Он содержит битекст с английского на другие 43 языка для 6193 предложений. Целью этих данных было предоставить стартовый набор чистых данных на разнообразные темы на этих языках.

• Как использовать данные. Вы можете получить доступ к ссылкам на данные в README по адресу https://github.com/facebookresearch/fairseq/tree/nllb

• Поддерживаемые задачи и рейтинги. Модель NLLB использует эти данные для улучшения производительности в языках с ограниченными ресурсами.

• Языки. NLLB-Seed содержит 43 языковых пар с английским.

Создание датасета

• Методика создания. Сценарий, диалект, орфография и подходы к переводу были установлены и согласованы на основе Flores-200. Переводчики использовали эти лингвистические выравнивания при работе над переводами NLLB-Seed. Датасеты были переведены непосредственно с английского на 39 языков, половина данных на лигурский язык (3000 предложений) сначала была переведена с английского на итальянский, затем с итальянского на лигурский, в то время как другая половина была переведена непосредственно с английского, и три арабских языка (Ачехский, Банджарский, Тамашек) были транслитерированы из соответствующих латинских датасетов, переведенных с английского. После фазы перевода или транслитерации была проведена фаза оценки лингвистического качества, в которой завершенные датасеты проверялись по лингвистическим выравниваниям из Flores-200, а также проводились базовые проверки качества. Датасеты были затем завершены и окончательно подготовлены.

• Исходные данные. Исходные данные включают 6193 английских предложения, отобранные из статей Википедии в 11 категориях: Антропология, Искусство, Биология, География, История, Математика, Люди, Философия, Физика, Общество, Технологии.

• Аннотации. Дополнительных аннотаций с битекстом нет.

• Личная и конфиденциальная информация. Не применимо

Рассмотрение использования данных

• Социальное воздействие датасета. Датасет специально создан для повышения качества перевода и улучшения идентификации языков в языках с крайне ограниченными ресурсами. Это помогает улучшить качество различных языков в системах машинного перевода.

• Обсуждение предвзятостей. Предвзятости в датасете не изучались.. Дополнительная информация

• Организаторы датасета. Все переводчики, участвовавшие в создании данных NLLB-Seed, прошли процесс проверки от наших партнёров-переводчиков. Переводчики должны быть носителями языка целевого текста и иметь высокий уровень владения английским языком (C1-C2). Для переводчиков, не являющихся носителями английского, требуется высокий уровень владения их родным языком. Также переводчики должны иметь не менее двух-трёх лет опыта в переводе в соответствующей языковой паре, если у них есть академическая степень по переводу или лингвистике, и три-пять лет опыта в переводе, если у них нет соответствующего академического образования. Переводчики также проходят тест перевода каждые 18 месяцев для оценки качества их переводов.

• Информация о лицензировании. Мы выпускаем переводы на основе исходных предложений из Википедии в соответствии с условиями лицензии CC-BY-SA.

• Информация о цитировании. Команда NLLB и др., No Language Left Behind: Scaling Human-Centered Machine Translation, Arxiv, 2022

Appendix H. Карточка данных для NLLB Multi-Domain данных

Описание датасета

• **Сводка датасета**. NLLB Multi-Domain представляет собой коллекцию человечески переведенных данных из четырёх доменов (11810 предложений из новостей, формальной и неформальной речи, и медицинских источников). Он содержит битекст с английского на шесть других языков. Целью этих данных является помощь в улучшении производительности модели на текстах из различных доменов и оценка того, насколько хорошо общая модель перевода может быть настроена на датасет, охватывающий новый домен.

• **Как использовать данные**. Вы можете получить доступ к ссылкам на данные в README по адресу https://github.com/facebookresearch/fairseq/tree/nllb

• **Поддерживаемые задачи и рейтинги**. Модель NLLB использует эти данные для улучшения производительности в языках с ограниченными ресурсами.

• **Языки**. NLLB Multi-Domain содержит 6 языковых пар с английским: Центральный аймара (ayr_Latn), Бходжпури (bho_Deva), Дюла (dyu_Latn), Фриульский (fur_Latn), Русский (rus_Cyrl) и Волоф (wol_Latn).

Создание датасета

• **Методика создания**. Сценарий, диалект, орфография и подходы к переводу были установлены и согласованы на основе Flores-200. Переводчики использовали эти лингвистические выравнивания при работе над переводами NLLB Multi-Domain. Датасеты были переведены непосредственно с английского на все шесть языков, после чего последовала фаза оценки лингвистического качества, в ходе которой завершённые датасеты проверялись по лингвистическим выравниваниям из Flores-200, а также проводились базовые проверки качества. Датасеты были затем завершены и окончательно подготовлены.

• **Исходные данные**. Исходные данные включают 3 домена:

– Новости: 2810 английских предложений из развивающегося набора данных WMT21 английско-немецкого языков, содержащего выборку газет 2020 года (Akhbardeh и др., 2021)

– Неформальная устная речь: 3000 английских высказываний из многосессионного чат-датасета Xu и др. (2022), включающего в среднем 23 слова на высказывание

– Здравоохранение: 3000 английских предложений из отчёта Всемирной организации здравоохранения (Donaldson и Rutter, 2017) и английской части отчёта TAUS о кризисе COVID-19

• **Аннотации**. Дополнительных аннотаций с битекстом нет.

• **Личная и конфиденциальная информация**. Не применимо. Рассмотрение использования данных

• **Социальное воздействие датасета**. Датасет специально создан для повышения качества перевода и идентификации языков в языках с крайне ограниченными ресурсами, что способствует улучшению качества различных языков в системах машинного перевода.

• **Обсуждение предвзятостей**. Предвзятости в датасете не изучались.

Дополнительная информация

• **Организаторы датасета**. Все переводчики, участвовавшие в создании данных NLLB Multi-Domain, прошли процесс проверки от наших партнёров-переводчиков. Переводчики должны быть носителями языка целевого текста и образованы в целевом языке. Они также должны обладать высоким уровнем владения английским языком (C1-C2). Для переводчиков, не являющихся носителями английского, требуется высокий уровень владения их исходным языком. Также переводчики должны иметь не менее двух-трёх лет опыта в переводе в соответствующей языковой паре, если у них есть академическая степень по переводу или лингвистике, и три-пять лет опыта в переводе, если у них нет соответствующего академического образования. Переводчики также проходят тест перевода каждые 18 месяцев для оценки качества своих переводов и для справки для всех будущих проектов.

• **Информация о лицензировании**. Мы выпускаем переводы на основе исходных предложений от Всемирной организации здравоохранения в соответствии с условиями лицензии CC-BY-SA. Мы также выпускаем переводы на основе исходных предложений от TAUS, Multi-Session Chat и WMT в соответствии с условиями лицензии CC-BY-NC.

• **Информация о цитировании**. Команда NLLB и др., No Language Left Behind: Scaling Human-Centered Machine Translation, Arxiv, 2022

Appendix I. Карточка данных для Метаданных добытого битекста

Описание датасета

• Сводка датасета

Мы создали добытый битекст из общедоступных веб-данных для 148 языковых пар с английским в центре и 1465 языковых пар без английского центра с использованием библиотеки stopes mining и кодировщиков LASER3 Heffernan и др. (2022). Мы открыли исходные данные метаданных и скрипт, позволяющий исследователям, скачавшим указанные файлы из CommonCrawl и ParaCrawl, воссоздать полные данные битекста. Обратите внимание, что CommonCrawl отвечает на уведомления о снятии материала, поэтому последующие запуски инструмента могут привести к меньшему объему битекста.

• Как использовать данные

Вы можете получить доступ к ссылкам на данные в README по адресу https://github.com/facebookresearch/fairseq/tree/nllb

Структура данных

• **Файлы метаданных** являются пробелами разделёнными, сжатыми файлами формата xz. Каждый файл соответствует одному направлению битекста. Например, файл xho_Latn-yor_Latn.meta.xz содержит все метаданные, необходимые для поиска фактических данных выравнивания текста на ксоса и йоруба. Каждая строка имеет 11 столбцов в следующем формате:

– Если метаданные извлечены из Common Crawl: wet_file_url document_sha1

document_url line_number_in_document paragraph_digest sentence_digest

lid_score laser_score direction language line_number_in_direction

– Если метаданные извлечены из другого корпуса: corpus_name.language not_used

not_used line_number_in_document paragraph_digest sentence_digest

lid_score laser_score direction language line_number_in_direction

• **Дайджесты параграфов и предложений** вычисляются с использованием xxh3_64_intdigest.

Разделение данных

• **Учитывая шумный характер всего процесса**, мы рекомендуем использовать данные только для обучения и использовать другие датасеты, такие как Flores-200, для оценки.

Создание датасета

• Исходные данные

Начальная коллекция данных и нормализация. Монолингвальные данные из Common Crawl и ParaCrawl.

• Обоснование кураторства

Мы применили фильтрацию на основе идентификации языка, фильтрации на основе эмодзи и для некоторых языков с высокими ресурсами - фильтрацию на основе языковых моделей. Дополнительные сведения о фильтрации данных можно найти в разделе 5.2.

• Кто являются источниками языковых продюсеров?

Исходный язык был создан авторами каждого веб-сайта, который был проиндексирован Common Crawl и ParaCrawl.

• Аннотации

– **Процесс аннотации**

Параллельные предложения в монолингвальных данных были идентифицированы с использованием кодировщиков LASER3 (Heffernan и др., 2022)

- Кто являются аннотаторами?

Данные не были аннотированы вручную.

- Личная и конфиденциальная информация

Файлы метаданных не содержат текста за пределами URL веб-сайтов. Однако данные из CommonCrawl и ParaCrawl могут содержать лично идентифицируемую информацию, чувствительное или токсичное содержимое, которое было публично опубликовано в интернете. Некоторая информация из этого может быть включена в выпущенный датасет.

Рассмотрения использования данных

• **Социальное воздействие датасета.** Эти данные могут быть использованы для воссоздания датасета для обучения систем машинного обучения на множестве языков с ограниченными ресурсами.

• **Обсуждение предвзятостей**. Предвзятости данных не были специально изучены, однако так как оригинальным источником данных является Всемирная паутина, вероятно, что данные обладают предвзятостями, схожими с теми, что распространены в интернете. Данные также могут содержать предвзятости, внесённые техниками идентификации языка и фильтрации данных: языки с ограниченными ресурсами, могут иметь более низкую точность, а техники фильтрации данных могут удалять некоторые менее естественные высказывания.

Дополнительная информация

• **Организаторы датасета**. Данные не были курированы.

• **Информация о лицензировании**. Мы выпускаем метаданные и скрипт для воссоздания битекста на их основе в соответствии с условиями лицензии CC-BY-NC. Текст и авторские права (при применимости) остаются у оригинальных авторов или издателей, пожалуйста, соблюдайте применимые лицензии, предоставленные оригинальными авторами. Мы отслеживаем исходный URL каждого индивидуального предложения, чтобы позволить людям ссылаться на указанный веб-сайт для получения информации о лицензировании.

• **Информация о цитировании**. Команда NLLB и др., No Language Left Behind: Scaling Human-Centered Machine Translation, Arxiv, 2022