

УДК 519.766

Разработка инструмента для автоматического извлечения параллельных предложений для языка малого народа России

Чапаев Артем Юрьевич

Козов Алексей Владимирович (научный руководитель)

МГТУ им. Н.Э. Баумана, Москва, Россия

Представлена система автоматического выравнивания русско-мокшанских текстов. Инструмент объединяет классификатор на основе искусственной нейронной сети (ИНС) с архитектурой fastText для определения языка и дообученную для выравнивания предложений ИНС с архитектурой «трансформер». Многошаговый итеративный цикл «ручная разметка → дообучение → поиск новых пар» позволил увеличить исходный корпус с 14 000 до 75 000 корректно выровненных предложений, при этом косинусное сходство параллельных пар выросло с 0,56 до 0,91. Полученные предложения опубликованы в открытый доступ и представляют собой крупный датасет параллельных предложений для мокшанского языка.

Ключевые слова: машинный перевод, выравнивание текстов, параллельные предложения, языки малых народов России, искусственная нейронная сеть

Введение. В мире существует около 7000 языков, однако высококачественные инструменты машинного перевода созданы менее чем для 250 из них. Мокшанский язык (ISO 639-3: mdf) — язык примерно 300 000 носителей — лишен открытого параллельного корпуса, что блокирует развитие машинного перевода. Машинный перевод позволяет сохранить культуру народов, обеспечить людей переводом из специфических доменов: медицина, технологии, право.

Задачи работы — разработать инструмент автоматического извлечения параллельных предложений для мокшанского и русского языков, собрать датасет параллельных предложений, достаточный по объему и качеству для дообучения инструмента машинного перевода.

Обзор существующих подходов. Существуют различные подходы к машинному переводу: лингвистический, статистический, нейросетевой. Сегодня наиболее распространенным является нейросетевой машинный перевод. Современные нейросетевые инструменты достигают близкого к человеческому качества при наличии не менее 10^6 пар предложений. При этом для малых языков объемы корпусов на порядки меньше.

Зачастую малые языки совсем не имеют корпусов предложений (например, как мокшанский), по этой причине важной задачей является сбор параллельных предложений. Для сбора параллельных предложений часто используется инструмент для определения языка (LangID) и инструмент поиска параллельных предложений.

Существующие идентификаторы языка (GlottLID-M [1], OpenLID [2]) не поддерживают мокшанский язык. Искусственные нейронные сети (ИНС)

LASER [3] и LaBSE [4], которые часто используются в инструментах для поиска параллельных предложений, не обучались на мокшанском корпусе.

Таким образом, задачи сбора данных и разработки инструмента автоматического извлечения параллельных предложений для мокшанского и русского языков являются важными для развития машинного перевода малых языков.

Постановка задачи. Рассмотрим математическую постановку задач извлечения параллельных предложений.

Задача определения языка: задано множество языков $L = \{l_1, l_2, \dots, l_K\}$, где K — общее число языков. Задача заключается в определении функции f такой, что $f: X \rightarrow L$, где X — множество текстов. Функция f должна возвращать для каждого объекта $x \in X$ вероятностное распределение $p(l \vee x)$ по всем классам $l \in L$.

Задача поиска параллельных предложений состоит из двух последовательных шагов: нахождение векторного представления предложения и выравнивание предложений.

Пусть X — множество предложений, а $E \subset R^d$ — пространство размерности d , тогда задача нахождения векторного представления состоит в определении функции g такой, что $g: X \rightarrow E$. Функция g принимает на вход объект $x \in X$ и возвращает векторное представление этого объекта $e \in E$.

Задача выравнивания между двумя множествами предложений S_1 и S_2 заключается в реализации алгоритма, который находит пары параллельных предложений (s_1, s_2) , где s_1 — предложение из множества S_1 , а s_2 — из множества S_2 . Алгоритм использует векторные представления всех предложений и находит такие предложения, что косинусная близость векторных представлений этих предложений максимальна. Другим свойством алгоритма является монотонность выравнивания: если S_1 и S_2 упорядочены, то не существует таких двух пар (s_1, s_2) и (s'_1, s'_2) , что если s_1 предшествует s'_1 , то s'_2 предшествует s_2 .

Разработка инструмента для извлечения параллельных предложений. Первым этапом разработки является подготовка данных для инструмента определения языка текста и для инструмента поиска параллельных предложений. В результате сбора данных для инструмента определения языка текста извлечено 118 000 статей Wikipedia на 341 языке. Также подготовлено 10 000 мокшанских и 10 000 эрзянских предложений из Библии для лучшего качества итогового инструмента. Эрзянский язык похож на мокшанский, поэтому важно точно различать эти близкие языки.

Для поиска параллельных предложений извлечено 14 000 пар параллельных предложений из Библии, словарей и учебников. Также извлечены тексты на мокшанском языке и их переводы на русский: 4 500 новостных статей, 3 000 статей из Wikipedia и 17 литературных произведений. Они использовались для дальнейшего итеративного увеличения выборки.

Сбор данных осуществлялся с помощью Python-модуль BeautifulSoup при работе с данными в формате HTML, а при работе с PDF — инструмент Tesseract.

Для разработки инструмента определения языка текста использована легковесная архитектура fastText [5]. После подбора гиперпараметров на валидационной выборке получена итоговая ИНС с высоким значением метрики F1-score, которая является средней гармонической от точности и полноты. Разработан класс LanguageDetector, решающий задачу определения языка текста.

Для разработки инструмента поиска параллельных предложений в качестве основы взята ИНС LaBSE-en-ru (128 млн параметров) [6]. Расширен словарь на 11 603 токена, которые являлись наиболее частыми парами токенов исходного словаря. Дальнейшее дообучение состояло из оптимизации сразу нескольких функций потерь. В рамках дообучения происходил итеративный процесс: ИНС дообучалась на имеющемся корпусе данных, производился поиск параллельных предложений среди извлеченных параллельных текстов, затем увеличивался корпус данных для дообучения для первого этапа. В результате через несколько итераций все параллельные тексты выровнены. Для более качественного итогового датасета произведен анализ пар с наименьшим значением косинусной близости. В ходе дообучения косинусное сходство для пар из валидационной выборки заметно увеличилось: для некоторых пар близость выросла с 0,56 до 0,91. Для выравнивания реализован алгоритм динамического программирования. Разработан класс SentenceAligner, реализующий поиск параллельных предложений в двух текстах на разных языках.

С использованием полученных инструментов получена итоговая выборка, которая включала около 75 000 параллельных предложений на мокшанском и русском языках. Часть параллельных предложений, которая извлечена из официального портала органов государственной власти Республики Мордовия, опубликована на HuggingFace. В этом датасете содержатся не только русско-мокшанские предложения, но и русско-эрзянские и мокшано-эрзянские, так как сайт содержит статьи в том числе на этом языке. Таким образом, с использованием итеративного подхода собран первый крупный датасет параллельных предложений на мокшанском языке.

Подбор гиперпараметров. При обучении fastText для задачи LangID произведен подбор оптимальных гиперпараметров. В результате получено значение метрики F1-score 0,82, что является достаточным для отбора текстов на мокшанском языке.

При дообучении ИНС для инструмента поиска параллельных предложений проверены различные гипотезы:

- 1) дообучение сразу всей ИНС;
- 2) первичное дообучение слоя преобразования токенов в векторные представления с дальнейшим дообучением всей ИНС;
- 3) первичное дообучение ИНС на параллельных словах с последующим дообучением на всех параллельных данных.

В результате наилучший результат на тестовой выборке получен с первичным дообучением слоя преобразования токенов в векторные представления на параллельных словах. Потом слой преобразования токенов в вектор-

ные представления дообучался на всех параллельных данных, а затем вся ИНС дообучалась на всех параллельных данных.

В табл. 1 указаны примеры из тестовой выборки. Косинусная близость при использовании недообученной LaBSE-en-ru для примеров из тестовой выборки отображена в табл. 2. После дообучения значение косинусной близости выросло, ее значения показаны в табл. 3.

Таблица 1

Соответствие русских и мокшанских предложений из тестовой выборки

Русский текст	Мокшанский текст
Картофель	Модамарь
Резать хлеб	Керемс кши
Поэтому, прежде всего, я бы хотел поздравить вас с профессиональным праздником.	Сяс, васендакиге, монь мялезе поздравляндамс тинь профессиональной илантень мархта.
Возле костра стоял большой, перепачканный сажей жестяной чайник.	Толнять тейса ащесь оцю соду жестень чайник.
Сидишь, положим, на возу, а ребята сдалька завидят: «Чапаев идет, Чапаев идет...»	Ащат озада, мярьктяма, усф лангса, а цјратне ичкозде нйясазь: «Чапаевсь сай, Чапаевсь сай...»

Таблица 2

Сходство между русскими и мокшанскими предложениями для необученной ИНС

	1	2	3	4	5
1	0,4047	0,3061	0,1994	0,0953	0,1105
2	0,3327	0,3384	0,0906	0,1487	0,1572
3	0,0122	0,1668	0,5606	0,1516	0,2414
4	0,0401	0,1449	-0,0137	0,2443	0,0003
5	0,0747	0,1684	0,1775	0,1386	0,6888

Таблица 3

Сходство между русскими и мокшанскими предложениями для итоговой ИНС

	1	2	3	4	5
1	0,7944	0,2908	-0,1055	-0,0320	-0,1035
2	0,4549	0,7198	0,0164	-0,0055	0,0235
3	0,0408	0,1490	0,8293	0,0158	0,0518
4	0,0097	-0,0553	-0,1475	0,5918	-0,0352
5	0,0433	-0,0256	-0,0863	-0,0318	0,6498

Заключение. В результате работы собран и опубликован корпус параллельных предложений для мокшанского языка объемом 75 000 пар. Разработан двухкомпонентный инструмент извлечения данных, превосходящий существующие решения. Используемое в работе итеративное дообучение LaBSE-en-ru показало эффективность при малом исходном корпусе.

Литература

- [1] Kargaran A.H. GlotLID: Language Identification for Low-Resource Languages. Findings of the Association for Computational Linguistics: EMNLP 2023. Singapore, Association for Computational Linguistics, 2023, pp. 6155–6218.
- [2] Burchell L. An Open Dataset and Model for Language Identification. Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics. Toronto, Association for Computational Linguistics, 2023, pp. 865–879.
- [3] Artetxe M., Schwenk H. Massively Multilingual Sentence Embeddings for Zero-Shot Cross-Lingual Transfer and Beyond. Transactions of the Association for Computational Linguistics. Cambridge, MA, 2019, vol. 7, pp. 597–610.
- [4] Feng F. Language-agnostic BERT Sentence Embedding. Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics. Dublin, Association for Computational Linguistics, 2022, pp. 878–891.
- [5] Bojanowski P. Enriching Word Vectors with Subword Information. Transactions of the Association for Computational Linguistics. Cambridge, MA, 2017, vol. 5, pp. 135–146.
- [6] LaBSE-en-ru: *модифицированная версия* LaBSE. URL: <https://habr.com/ru/articles/562064> (accessed 14.03.2025).

Development of a tool for automatic extraction of parallel sentences for the language of the small nation of Russia

Chapaev Artem Yurievich

Kozov Alexey Vladimirovich (*scientific advisor*)

BMSTU, Moscow, Russia

The paper presents a system for automatic alignment of Russian–Moksha texts. The tool combines a fastText-based neural classifier for language identification with a Transformer-based neural network fine-tuned for sentence alignment. A multistage iterative cycle of “manual annotation → fine-tuning → search for new pairs” expanded the initial corpus from 14,000 to 75,000 correctly aligned sentences, while the cosine similarity of parallel pairs rose from 0.56 to 0.91. The resulting sentences have been released as open data and constitute a large dataset of parallel sentences for the Moksha language.

Keywords: machine translation, text alignment, parallel sentences, minority languages of Russia, artificial neural network