**Отчет**

о результатах исследования по использованию

алгоритмов машинного обучения для перевода текста

**Цель исследования:** разработка моделей с программным интерфейсом для машинного перевода исходного текста в целевой с учетом предметной области, определяемой заказчиком продукта.

**Задачи исследования:**

1. Изучить существующие подходы к построению алгоритмов машинного перевода с целью их программной реализации.
2. Провести сравнительный анализ существующих моделей машинного перевода для выбора наиболее применимых в разработке программного интерфейса.
3. Сформировать перечень источников наборов данных для обучения моделей машинного перевода.
4. Определить аппаратное обеспечение для применения алгоритмов машинного обучения моделей при обработке естественного языка.
5. Разработать управляемый механизм разработки моделей машинного перевода, который позволит в целевом языке учитывать предметную область заказчика.
6. Разработка программного интерфейса для машинного перевода исходного текста в целевой

# 1. Существующие подходы к построению алгоритмов машинного перевода

Существует несколько принципиально разных подходов к построению алгоритмов машинного перевода: основанный на правилах (rule-based), статистический, или основанный на статистике (statistical-based), нейронный машинный перевод (neural machine translation, NMT). Последний подход получил наибольшее развитие благодаря развитию современных архитектур нейронных сетей и распространению решений с открытым исходным кодом.

В открытом доступе существует значительное число ресурсов с готовыми решениями нейронного машинного перевода. Наиболее популярными решениями являются:

|  |  |  |  |
| --- | --- | --- | --- |
| **Решение** | **Язык** | **Фреймворк** | **Число языков** |
| Tensor2Tensor | Python | TensorFlow |  |
| FairSeq | Python | PyTorch |  |
| Nmt | Python | TensorFlow |  |
| OpenNMT | Python/C++ | PyTorch/TensorFlow |  |
| Sockeye | Python | MXNet |  |
| Nematus | Python | Tensorflow |  |
| Marian | C++ | – | 177 |
| THUMT | Python | PyTorch/TensorFlow |  |
| NMT-Keras | Python | Keras |  |
| Neural Monkey | Python | TensorFlow |  |
| Transformers Hub | Python | PyTorch/TensorFlow/JAX | 177 |

В ходе проведенного исследования были проведены эксперименты с Marian для машинного перевода.

Marian написан на чистом C ++ [1]. Основной зависимостью Marian является Boost. Marian может быть скомпилирован на машинах с устройствами NVIDIA GPU и CUDA 8.0+ или на машинах с CPU. Версия Marian для ЦП компилируется автоматически, если обнаружены OpenBLAS или Intel MKL (рекомендуется).

Эксперименты показали, что полноценное использование моделей для Marian можно свободно реализовать под OS Linux. Однако в OS Windows могут возникнуть проблемы, связанные с выполнением дополнительных скриптов по предобработке исходного текста; без такой предобработки может быть получен перевод, который меняет смысл исходного текста (например, «документом о Кремлях» вместо корректной фразы «документом о взглядах Кремля»; «встретить репрессивную политику государства» вместо корректной фразы «противостоять репрессивной политике государства»).

Кроме этого, при использовании Marian требуется выполнять компиляцию под целевую операционную систему, что снижает гибкость его использования, в т.ч. при обучении моделей. Так, для одной эпохи при обучении на центральном процессоре (Intel(R) Core(TM) i9-9900 CPU @ 3.10GHz, 3096 МГц, ядер: 8, логических процессоров: 16) с корпусом из 4468840 парных предложений ушло 27 часов 44 минуты.

Дальнейшие исследования определили поиск более гибких решений, среди которых следует выделить [Transformers](https://pypi.org/project/transformers/), которое предоставляет доступ к набору данных, словарям и предварительно обученным моделям машинного перевода (размещены на сайте [huggingface.co](https://huggingface.co/)), их дообучения и локального использования. Экосистема Transformers включает в себя набор инструментов для работы с набором данных, распределенного использования аппаратного обеспечения для машинного обучения и оценки моделей (Transformers, Datasets, Tokenizers и Accelerate).

Сравнительный анализ Marian c Transformers в скорости перевода [тестового набора](https://github.com/eleldar/Translator/raw/master/test_dataset/test_dataset.xlsx) показал следующие результаты (Intel(R) Core(TM) i9-9900 CPU @ 3.10GHz, 3096 МГц, ядер: 8, логических процессоров: 16):

|  |  |  |
| --- | --- | --- |
|  | **Англо-русский** | **Арабско-русский** |
| Marian | 99,1 сек. | 94,6 сек. |
| Transformers | 183,4 сек. | 152,7 |
| Разница (абс/отн) | 84,3 / 1,8 | 58.1 / 1,6 |

Из таблицы видно, что Marian работает значительно быстрее. Однако отмеченные выше ограничения на использование его моделей для обучения и практического применения определил Transformers в качестве более перспективной основы для решения дальнейших задач исследования.

# 2. Сравнительный анализ существующих моделей машинного перевода

Существующие модели машинного перевода можно разделить по следующим критериям:

* архитектура модели;
* гиперпараметры, использованные при обучении;
* набор данных, использованный при обучении;
* метод, использованный при формировании словаря (токенизация);
* число языков, которые может обработать модель.

Основная часть современных моделей основана на трансформерах [2, 3, 4, 5]. Другие модели, как правило, используются в исследованиях, чтобы их сравнить, в том числе с трансформерами. В научных работах можно встретить тестирование таких моделей как amun и s2s [6, 7].

Основные архитектуры современных моделей машинного перевода:

|  |  |  |  |
| --- | --- | --- | --- |
| **Модель** | **Кодер** | **Декодер** | **Сложность при обучении** |
| [RCTM 1 (Kalchbrenner and Blunsom, 2013)](https://www.sciencedirect.com/science/article/pii/S2666651020300024#bib64) | CNN | RNN | O(S2+T) |
| [RCTM 2 (Kalchbrenner and Blunsom, 2013)](https://www.sciencedirect.com/science/article/pii/S2666651020300024" \l "bib64) | CNN | RNN | O(S2+T) |
| RNNEncdec/Seq2Seq (Cho et al., 2014a; Sutskever et al., 2014) | RNN | RNN | O(S+T) |
| [RNNsearch (Bahdanau et al., 2015)](https://www.sciencedirect.com/science/article/pii/S2666651020300024#bib10) | RNN | RNN | O(ST) |
| [ByteNet (Kalchbrenner et al., 2016)](https://www.sciencedirect.com/science/article/pii/S2666651020300024" \l "bib65) | CNN | CNN | O(S+T) |
| [ConvSeq2Seq (Gehring et al., 2017)](https://www.sciencedirect.com/science/article/pii/S2666651020300024" \l "bib43) | CNN | CNN | O(ST) |
| [Transformer (Vaswani et al., 2017)](https://www.sciencedirect.com/science/article/pii/S2666651020300024" \l "bib126) | SAN | SAN | O(S2+ST+T2) |

где RNN - recurrent neural network, CNN - convolution neural network, SAN - and self-attention network, S – длина слов исходного языка, T - длина слов целевого языка

На сайте [huggingface.co](https://huggingface.co/) размещено 1444 моделей машинного перевода с архитектурой Transformer. В разрезе моделей наибольшее число принадлежит следующим:

|  |  |
| --- | --- |
| **Модель** | **Количество** |
| marian | 1367 |
| T5 | 20 |
| mbart | 11 |

Исходя из этого в качестве предобученных моделей были взяты модели marian. Исследования показали, что среди этих моделей сейчас доступно 1339 языковых пар, из которых:

* уникальные языковые пары – 1330, повторы – 9.
* симметричные пары – 469 (всего 938 моделей).
* ассиметричные – 194.
* мультиязычные – 198.

Полный список доступных моделей сведен в [таблицу](https://github.com/eleldar/Translator/raw/master/table.xlsx). При этом их количество и содержание не статично, поскольку могут добавляться новые либо обновляться существующие.

# 3. Набор данных для обучения моделей машинного перевода

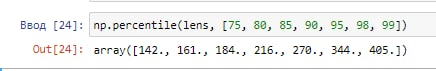
Из открытого доступа были загружены следующие корпуса:

- [англо-русского языков](https://object.pouta.csc.fi/Tatoeba-Challenge-v2021-08-07/eng-rus.tar) (217 612 504 предложений);

- [арабско-русского языков](https://object.pouta.csc.fi/Tatoeba-Challenge-v2021-08-07/ara-rus.tar) (60 875 366 предложений);

- [арабско-английского языков](https://object.pouta.csc.fi/Tatoeba-Challenge-v2021-08-07/ara-eng.tar) (126 590 524 предложений).

Для работы было сформировано 3 файла в формате CSV с длиной 405, не превышающие 405 символа. Данная длина выбрана на основе ограничений, которые вводятся на длину входных данных при обучении и описательной статистики для данных корпусов:



Для применения сформированных наборов данных в обучении моделей были созданы объекты с использованием модуля [memmap](https://numpy.org/doc/stable/reference/generated/numpy.memmap.html) библиотеки NumPy. Однако итоговый объем данных составил более 900 Гб, что вызвано особенностями организации хранения данных этой библиотекой.

С целью применения более эффективных решений по организации хранения и использования набора данных в настоящее время проводится изучение двух основных подходов: Apache Arrow (рис. 1) и генераторы потока (рис. 2).

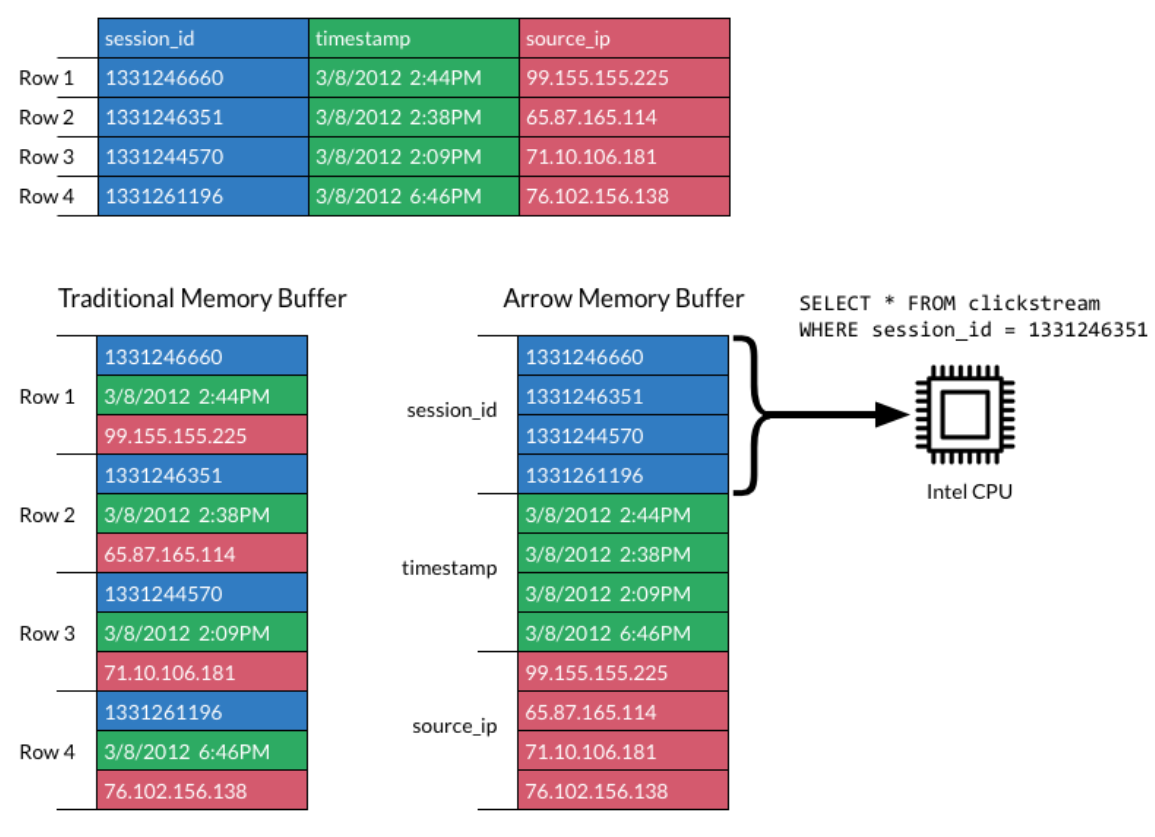


Рис. 1. Сравнение Apache Arrow с традиционным подходом к хранению данных

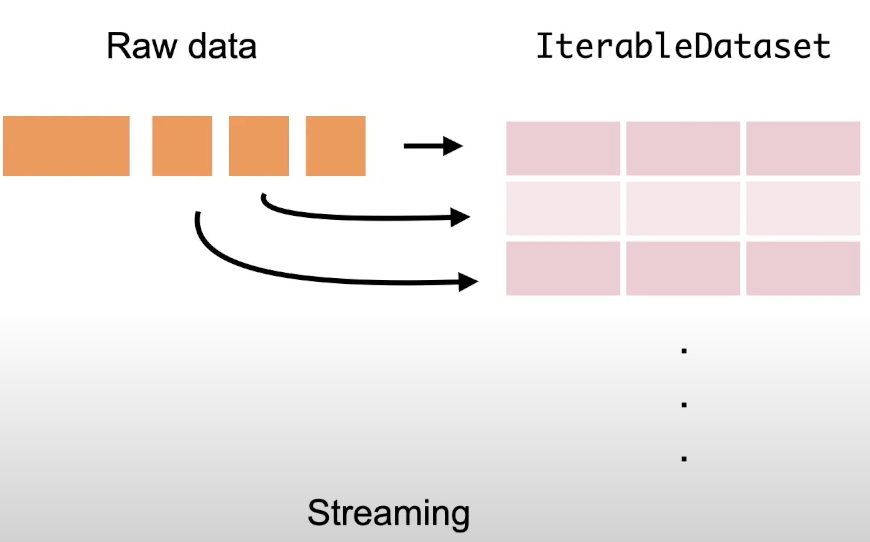


Рис. 2. Схема создания генератора потока набора данных

Оба подхода не требуют создания новых файлов, однако имеют свои преимущества и недостатки. Так, для создания объекта в оперативной памяти из корпуса 217 610 088 англо-русских предложений Apache Arrow потребовал 181 Мб, для создания генератора потока потребовалось всего 9 Мб. Вместе с тем, скорость обращения к элементам первого объекта в разы быстрее, чем ко второму.

С генератором потока набора данных был проведен эксперимент по подсчёту числа элементов, состоящий из 210 607 728 текстов (в архиве занимает 425 ГБ). Обычные циклы не справились с этой задачей в течение 24 часов, поэтому было применено распределенное вычисление на 16 логических процессорах, которое заняло 7 часов 15 минут.

Исходя из этого Apache Arrow был выбран в качестве основного инструмента для работы с данными. Он позволил не создавать дополнительные файловые объекты.

Для анализа и оценки результатов обучения моделей машинного перевода были сформированы 2 набора данных:

- англо-русский разговорник обыденных фраз из 2664 предложений;

- англо-русский разговорник в IT-сфере из 1038 предложений.

# 4. Аппаратное обеспечение для применения алгоритмов машинного обучения моделей при обработке естественного языка

Для выбора оптимального аппаратного обеспечения для обучения моделей машинного обучения на учебных данных был проведен эксперимент в Google Colab Pro, который показал следующие результаты. Для обработки корпуса из 1377 парных предложений потребовалось:

- CPU - 9 часов 15 минут;

- GPU (Tesla P100-PCIE) - 10 минут 16 секунд;

- TPU - 1 час 22 минуты.

Выбор Google Colab обусловлен тем, что для обучения в Kaggle не хватает памяти при выполнении алгоритма по обучению модели:



Рис. 3. Окно ошибки Kaggle

Аналогичная проблема возникает при попытке применить библиотеку для распределения ресурсов:



Рис. 4. Окно ошибки Kaggle

Аналогичный эксперимент проведен на локальном сервере с одним GPU (GeForce RTX 2080Ti). Скорость работы по сравнению c Google Colab Pro составила 7 минут 7 секунд.

Также следует отметить, что ключевой особенностью алгоритмов по обработке естественных языков является то, что они могут передавать различные размеры пакетов в модель. Эта особенность не позволяет реализовать в полной мере TPU для обучения моделей машинного обучения.

Исходя из этого, оптимальным выбором аппаратного обеспечения для обучения моделей машинного перевода является GPU.

# 5. Управляемый механизм разработки моделей машинного перевода

Создание управляемого механизм разработки моделей машинного перевода можно разбить на три основных направления:

1. Разработка абсолютно новых словаря и токенизатора с последующим обучением модели, веса которой были инициализированы случайными значениями;
2. Использование существующего токенизатора с последующим обучением модели, веса которой были инициализированы случайными значениями;
3. Использование существующего токенизатора с последующим обучением предобученной модели.

В рамках последнего направления был проведен эксперимент и в существующей модели удалось добиться желаемого результата.

Для обучения использовалось 79916 парных предложений, которые применены к модели 'Helsinki-NLP/opus-mt-en-ru'. Обучение проходило в течение 2-х эпох на GPU (Tesla P100-PCIE) на протяжении 2:37 часов.

Полученные результаты:

|  |  |  |
| --- | --- | --- |
| **Исходный текст** | **До обучения** | **После обучения** |
| Security of online payments | Обеспечение онлайновых платежей | Безопасность онлайн-платежей |

При сравнении словарей токенизаторов в marian на С++ с marian на Python выяснилось, что они одинаковые (отличие лишь в служебном символе Transformers). Поэтому для дообучения на этом токенизаторе придется выполнять предобработку либо строить свой токенизатор (но в этом случае дообучение будет равнозначно обучению с нуля, т.к. IDs вероятнее всего будут другими).

# 6. Программный интерфейс для машинного перевода

Для реализации программного интерфейса был разработан RESTful API на основе библиотек Python (flask и flask\_restful) с документированием спецификации по стандарту OpenAPI. В настоящее время документированный сервис доступен по адресу: <http://10.7.102.48:8888/api/>. Запросы прослушиваются по адресу: <http://10.7.102.48:8888/api/translation>.

Для работы с интерфейсом требуется передать 3 аргумента: исходный текст, код исходного языка и код целевого языка (например: "text": "Hello World!", "sourceLanguage": "en", "targetLanguage": "ru"). В ответ API возвращает этот запрос с ключом "message", содержащий переведенный текст.

Настоящее время доступны следующие направления перевода: en-ru, ru-en, ar-ru, ru-ar, en-ar, ar-en, где "en" – Английский, "ru" – Русский, "ar" – Арабский.

Текущая реализация сервиса была развернута в отдельном окружении (<http://10.239.16.89:60001/api/translation>) для сравнения эффективности работы CPU (Intel(R) Xeon(R) CPU E5-2630 v4 @ 2.20GHz) и GPU (GeForce RTX 2080Ti, 11019MiB) при переводе английского (таблица 1) и арабского (таблица 2) текстов на русский из [тестового набора](https://github.com/eleldar/Translator/raw/master/test_dataset/test_dataset.xlsx).

Таблица 1. Перевод английских предложений на русский язык

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Англо-русский перевод** | **Всего**  **(68 предложений)** | | | **Среднее значение**  **(на 1 предложение)** | |
| **10 CPU, 16 RAM**  **(доля)** | **1 GPU, 16 RAM**  **(доля)** | **Разница**  **(относит.)** | **10 CPU,**  **16 RAM** | **1 GPU,**  **16 RAM** |
| Предобработка | 2,71 мс  (0,002%) | 2,36 мс  (0,009%) | -0,35 мс  (-12,91%) | 0,04 мс | 0,035 мс |
| Токенизация | 52,55 мс  (0,048%) | 42,40 мс  (0,16%) | -10,15 мс  (-19,31%) | 0,77 мс | 0,62 мс |
| Генерация | 107727,76 мс  (99,79%) | 25704,70 мс  (99,52%) | -82023 мс  (-76,13%) | 1584,23 мс | 378,01 мс |
| Декодирование | 168,35 мс  (0,15%) | 77,69 мс  (0,30%) | -90,66 мс  (-53,85%) | 2,47 мс | 1,145 мс |
| **Всего:** | 107951,31 мс | 25827,15 мс | -82124,16 мс  (-76,07%) | 1587,51 мс | **379,81 мс** |
| Использование RAM | - | | | 16,06 Мб | 0,14 Мб |

Таблица 2. Перевод арабских предложений на русский язык

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Арабско-русский перевод** | **Всего**  **(57 предложений)** | | | **Среднее значение**  **(на 1 предложение)** | |
| **10 CPU, 16 RAM**  **(доля)** | **1 GPU, 16 RAM**  **(доля)** | **Разница**  **(относит.)** | **10 CPU,**  **16 RAM** | **1 GPU,**  **16 RAM** |
| Предобработка | 2,88 мс  (0,002%) | 2,67 мс  (0,01%) | -0,21 мс  (-7,29%) | 0,05 мс | 0,046 мс |
| Токенизация | 52,36 мс  (0,05%) | 41,75 мс  (0,17%) | -10,61 мс  (-20,26%) | 0,91 мс | 0,73 мс |
| Генерация | 102097,92 мс  (99,80%) | 24005,08 мс  (99,52%) | -78092,8 мс  (-76,48%) | 1791,19 мс | 421,14 мс |
| Декодирование | 148,95 мс  (0,14%) | 70,49 мс  (0,29%) | -78,46 мс  (-52,67%) | 2,61 мс | 1,23 мс |
| **Всего:** | 102302,11 мс | 24119,99 мс | -78182,1 мс  (-76,42%) | 1794,76 мс | **423,14 мс** |
| Использование RAM | - | | | 19,43 Мб | 0,16 Мб |

Далее был проведен эксперимент по дискретному изменению числа доступных CPU (Intel(R) Xeon(R) CPU E5-2630 v4 @ 2.20GHz, 16 Gb RAM) в виртуальной среде. Были рассмотрены вариации с 1, 2, 3, 4, 5, 10, 15, 20, 25 и 30 доступных CPU. Результаты эксперимента показали, что с увеличением числа доступных CPU для перевода английских (таблица 3) и арабских предложений (таблица 4) требуется меньше времени.

Таблица 3. Среднее значение потребляемых ресурсов

при переводе английских предложений

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **CPU** | **Предобработка,**  **мс** | **Токенизация,**  **мс** | **Генерация,**  **мс** | **Декодирование,**  **мс** | **Всего,**  **мс** | **Память**  **(Мб)** |
| **1** | 0,10 | 1,90 | 36755,19 | 59,13 | 36816,31 | 14,44 |
| **2** | 0,09 | 1,64 | 15872,11 | 18,23 | 15892,07 | 15,80 |
| **3** | 0,07 | 1,48 | 8963,24 | 9,43 | 8974,22 | 14,93 |
| **4** | 0,06 | 1,05 | 4896,26 | 16,49 | 4913,86 | 13,62 |
| **5** | 0,05 | 0,98 | 3983,98 | 7,30 | 3992,30 | 15,38 |
| **10** | 0,04 | 0,74 | 1593,74 | 4,09 | 1598,62 | 15,79 |
| **15** | 0,04 | 0,72 | 964,77 | 1,93 | 967,46 | 15,95 |
| **20** | 0,04 | 0,68 | 672,53 | 1,67 | 674,92 | 16,15 |
| **25** | 0,04 | 0,78 | 670,89 | 1,76 | 673,46 | 16,87 |
| **30** | 0,04 | 0,68 | 668,66 | 1,72 | 671,10 | 18,24 |

Таблица 4. Среднее значение потребляемых ресурсов

при переводе арабских предложений

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **CPU** | **Предобработка,**  **мс** | **Токенизация,**  **мс** | **Генерация,**  **мс** | **Декодирование,**  **мс** | **Всего,**  **мс** | **Память**  **(Мб)** |
| **1** | 0,12 | 3,74 | 41257,42 | 59,96 | 41321,23 | 19,63 |
| **2** | 0,09 | 2,02 | 17970,40 | 19,67 | 17992,17 | 16,28 |
| **3** | 0,07 | 5,10 | 10057,73 | 6,76 | 10069,67 | 18,57 |
| **4** | 0,06 | 1,23 | 5682,44 | 13,84 | 5697,57 | 17,12 |
| **5** | 0,06 | 1,17 | 4529,40 | 7,98 | 4538,60 | 20,56 |
| **10** | 0,05 | 0,92 | 1863,91 | 3,06 | 1867,94 | 19,25 |
| **15** | 0,04 | 0,82 | 1093,27 | 2,61 | 1096,73 | 15,21 |
| **20** | 0,04 | 0,85 | 758,21 | 1,82 | 760,92 | 20,10 |
| **25** | 0,04 | 0,82 | 756,53 | 1,95 | 759,34 | 18,62 |
| **30** | 0,04 | 0,84 | 769,44 | 1,96 | 772,29 | 18,30 |

Однако эта зависимость не линейна и для обоих языков при 20 CPU отмечается производительность, которая дальше не улучшается (рисунок 5).

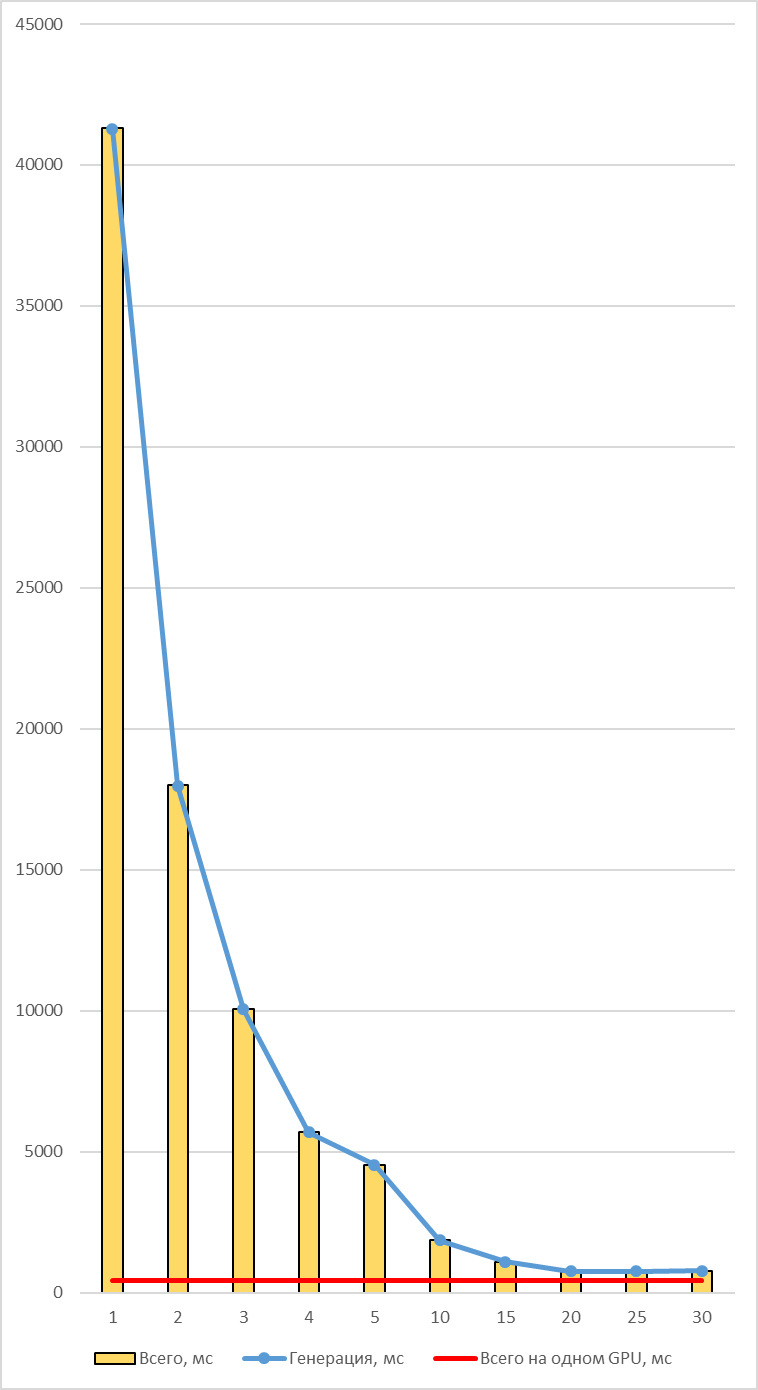
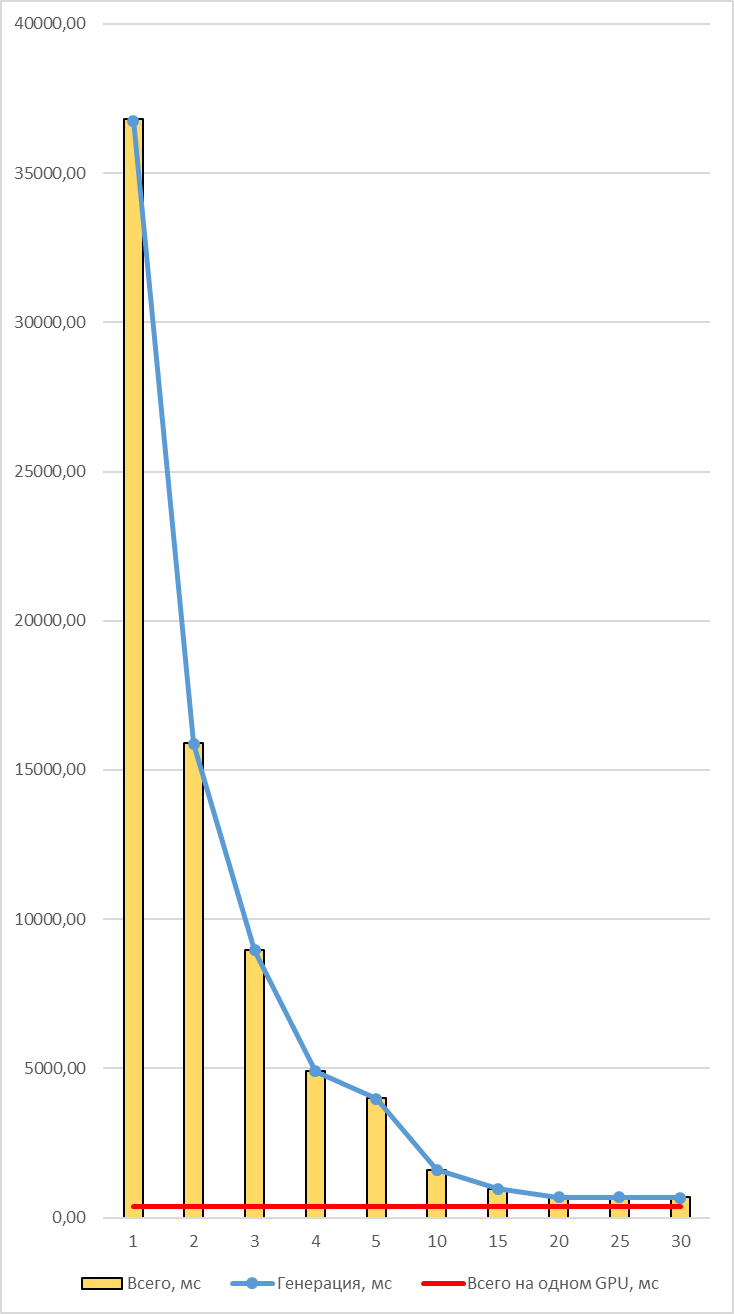


Рис. 5. Результаты перевода английских (слева) и арабских (справа)

Полученные результаты свидетельствуют о том, что производительность CPU асимптотически может приближаться к результатам работы сервиса перевода на GPU, но имеет свой предел, который значительно уступает последнему. Кроме того, GPU обладает возможностью пакетной обработки данных, что позволяет сократить время токенизации и генерации на размер пакета (например, при получении текста из нескольких предложений они разбиваются на отдельные предложения и формируются в отдельный пакет, далее этот пакет будет обработан токенизатором и моделью, т.е. несколько предложений будут обработаны одновременно).

Сравнить с работой Marian точно не получается, поскольку требуется сборка с реализацией функций по измерению каждого этапа. Произвольные замеры потребляемой памяти с момента запуска цикла перевода в Linux показали, что она колеблется от 20 до 60 Мб.

# Список использованных источников

1. [Marian: Fast Neural Machine Translation in C++ / Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak et al. // Proceedings of ACL 2018, System Demonstrations. — Melbourne, Australia: Association for Computational Linguistics, 2018. — July. — Pp. 116–121](https://aclanthology.org/P18-4020.pdf). http://www.aclweb.org/anthology/P18-4020.

2. [From Research to Production and Back: Ludicrously Fast Neural Machine Translation / Young Jin Kim, Marcin Junczys-Dowmunt, Hany Hassan et al. — 2019.](https://github.com/eleldar/Translator/blob/master/articles/d19_5632.pdf)

3. [Molchanov, Alexander. PROMT Systems for WMT 2018 Shared Translation Task / Alexander Molchanov. — 2019](https://github.com/eleldar/Translator/blob/master/articles/w18_6420.pdf).

4. [The AFRL IWSLT 2020 Systems: Work-From-Home Edition / Brian Ore, Eric Hansen, Tim Anderson, Jeremy Gwinnup. — 2020](https://github.com/eleldar/Translator/blob/master/articles/2020iwslt_111.pdf).

5. [Translating Between Morphologically Rich Languages: An Arabicto-Turkish Machine Translation System / ˙Ilknur Durgar El-Kahlout, Emre Bekta¸s, Naime ¸Seyma Erdem, Hamza Kaya. — 2019](https://github.com/eleldar/Translator/blob/master/articles/w19_4617.pdf).

6. [Marian: Cost-effective High-Quality Neural Machine Translation in C++ / Marcin Junczys-Dowmunt, Kenneth Heafield, Hieu Hoang et al. — 2019](https://github.com/eleldar/Translator/blob/master/articles/w18_2716.pdf).

7. [Tambouratzis, George. Alignment verification to improve NMT translation towards highly inflectional languages with limited resources / George Tambouratzis, Marina Vassiliou. — 2021](https://github.com/eleldar/Translator/blob/master/articles/2021eacl_main158.pdf).