Технологии подготовки обучающих данных для системы нейросетевого машинного перевода

А. Хусаинов, А. Гатиатуллин, Н. Прокопьев
Институт прикладной семиотики АН РТ

План доклада

- Основная идея проекта
- Результаты по созданию датасетов
- Результаты экспериментов
- Выводы

TurkLang - 7

Цель:

- обеспечить русско-тюркские языковые пары качественным переводчиком.

Задачи:

- **создание датасетов**;
- эксперименты с технологиями;
- построение нейросетевых моделей;
- ✓ создание веб-сайта – turk.translate.tatar;
- объединение усилий научных групп, энтузиастов.

TurkLang - 7

Цель:

- обеспечить русско-тюркские языковые пары качественным переводчиком.

Актуальность:

- сохранение и развитие языков;
- активизация использования языков в Интернете;
- возможность качественного перевода документов;
- изучение языков.

TurkLang - 7

Цель:

- обеспечить русско-тюркские языковые пары качественным переводчиком.

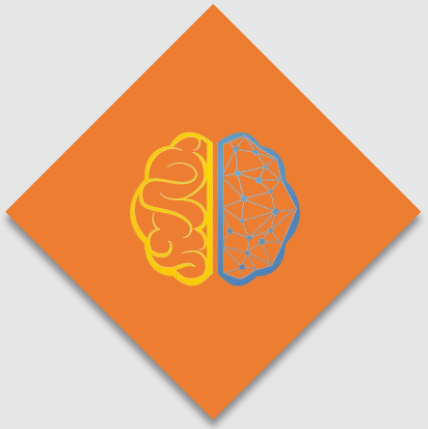
Татарский **Башкирский** **Казахский** **Чувашский** **Узбекский** **Киргизский** **Крымско-татарский**

58 миллионов носителей*

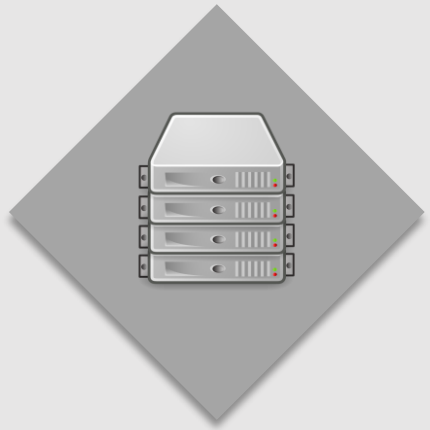
*согласно данным проекта Ethnologue
https://www.ethnologue.com/

Машинный перевод

ДАННЫЕ



ТЕХНОЛОГИИ



ВЫЧИСЛИТЕЛЬНЫЕ МОЩНОСТИ

При наличии всех трех составляющих

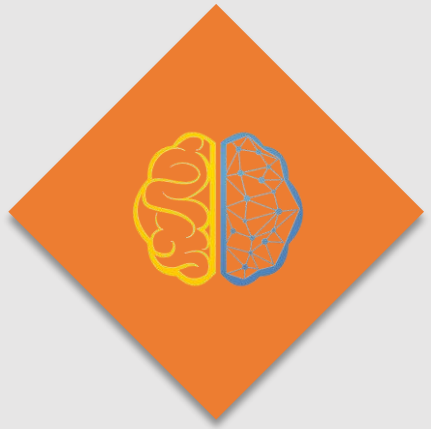
возможно получение хороших результатов

Основные этапы

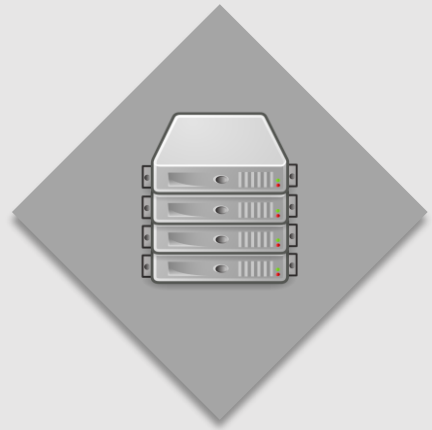
Подготовки корпусов

Машинный перевод

1. Поиск, обработка и выравнивание двуязычного контента
+ готовые корпуса



3. Back-translation

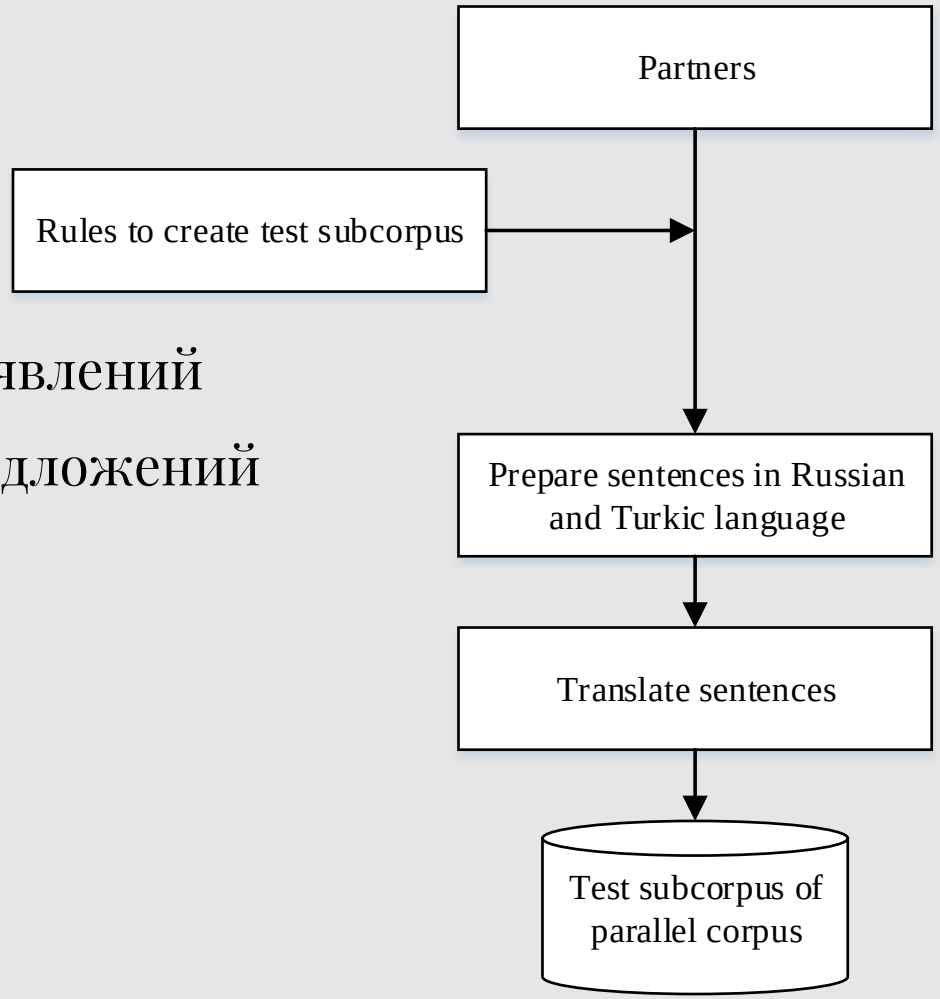


2. С помощью модели Тюркской морфемы

* Использование больших языковых моделей

TurkLang – 7. 1 путь

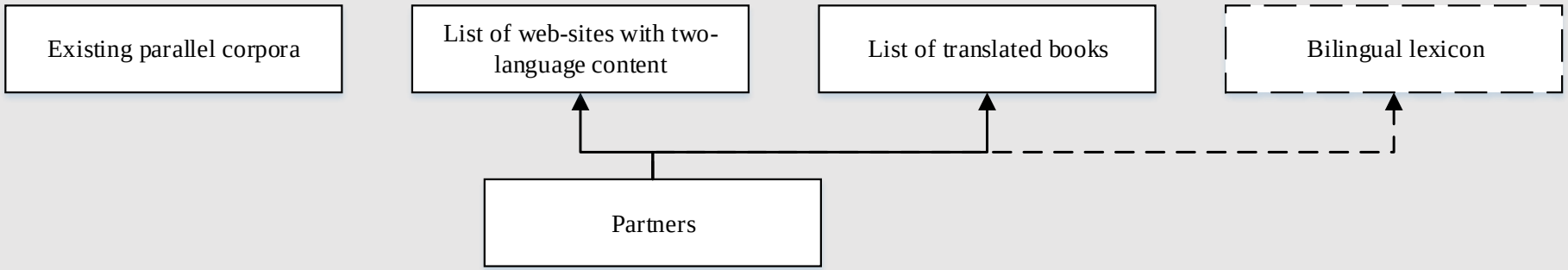
1. **Создание тестовых корпусов**
   - Проблема высоких BLEU
   - Экспертное описание языковых явлений
   - Полуавтоматический подбор предложений по каждому из явлений



TurkLang – 7. 1 путь

2. Источники двуязычных текстов

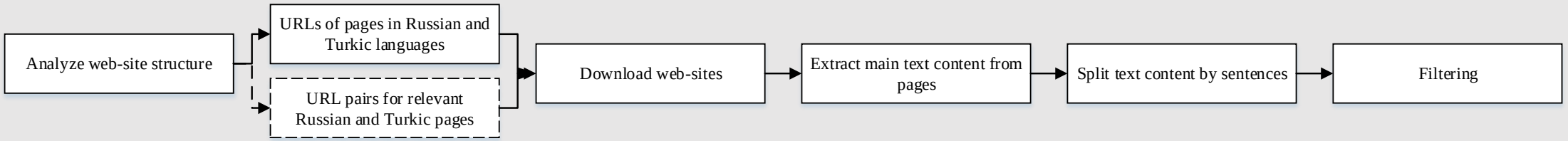
- Для большинства языков (KY, BA, TT, UZ) основной источник – веб-сайты
- Для крымско-татарского – книги
- Для казахского и чувашского – существующие открытые и закрытые датасеты



TurkLang – 7. 1 путь

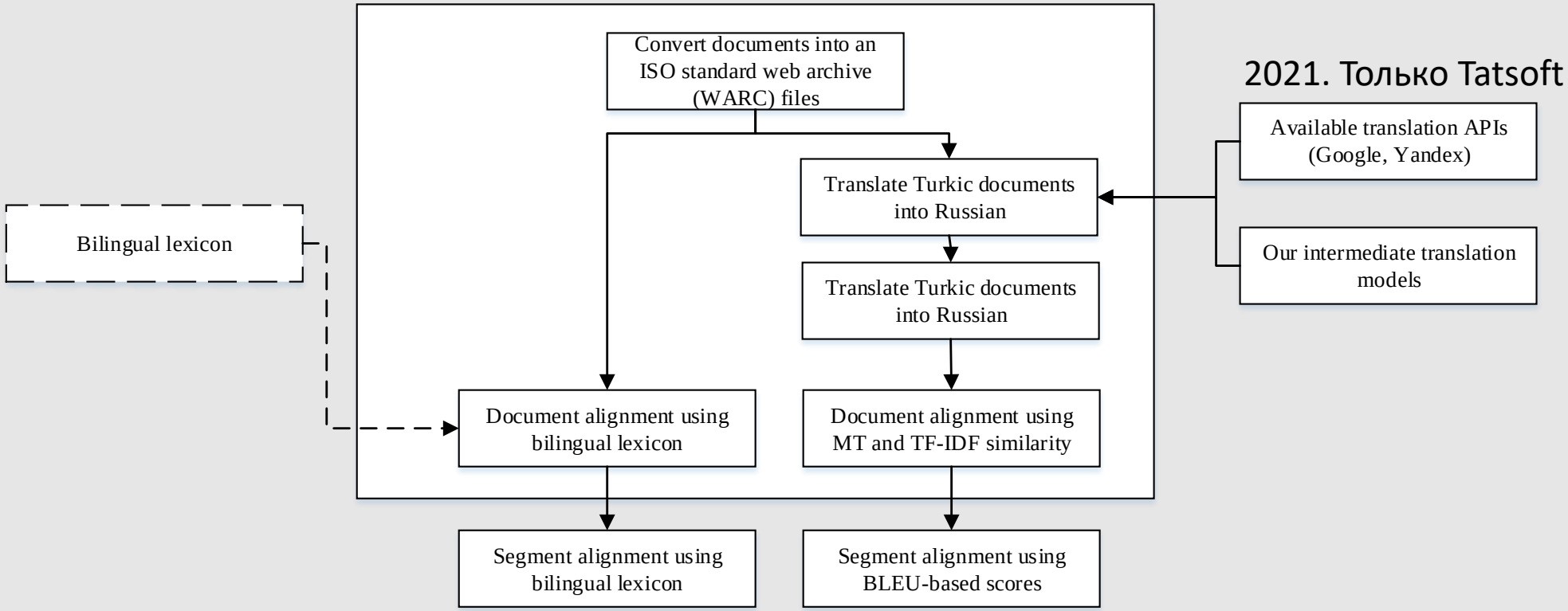
3. Сбор веб-данных

- Ручная работа:
  - формирование начального списка доменов
  - анализ структуры сайтов, проверка наличия файла Sitemap
  - оценка возможности точной установки пар переводных документов
- Trafilatura (+ отдельный скрипт для сайта минюста Киргизии)
- Разделение на предложения (https://github.com/natasha/razdel)

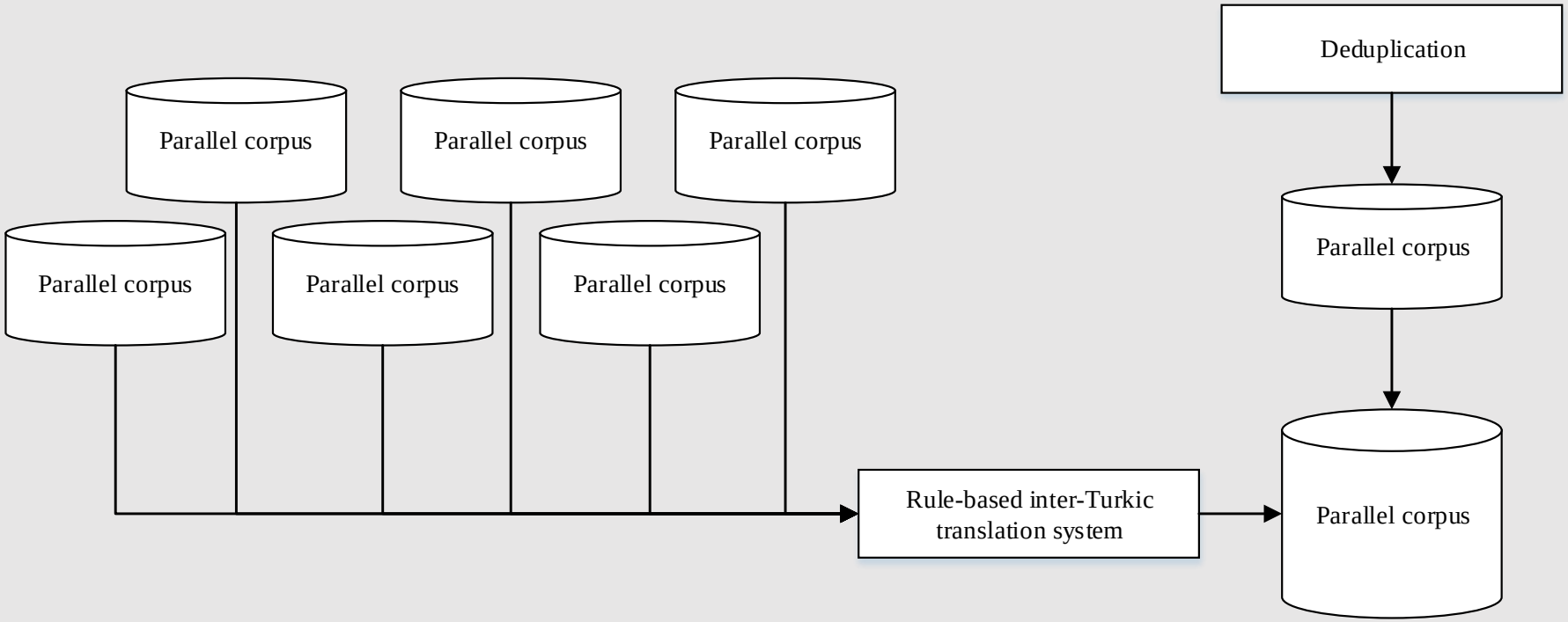


TurkLang – 7. 1 путь

4. Выравнивание документов и сегментов



TurkLang – 7. 2 путь



TurkLang – 7. 2 путь

Rule-based подход для межтюркского перевода

На базе портала тюркской морфемы - http://modmorph.turklang.net/

Описание тюркских языков на базе общего подхода:
- морфологический;
- синтаксический;
- семантический уровни.

Алгоритм перевода между тюркскими языками.
1. Морфологический разбор синтаксических конструкций и отдельных слов;
2. **Поиск и перевод «концептов»-основ слов;** **А если нет?**
3. Морфологический синтез слов и фраз на целевом языке.

Проблемы – наполненность баз данных портала, многозначность.

TurkLang – 7. 2 путь

Процесс «унификации» русско-тюркских корпусов:

0. Сбор моноязычных корпусов и построения ЯМ:
    - Части параллельных корпусов
    - Моноязычные порталы
    - Существующие корпуса
1. Варианты перевода, если слова нет в базе:
    - Оставить без изменений в переведённом предложении;
    - Безосновный переводчик – все вариант разборов;
2. Если слово имеет многозначный разбор:
    - Наиболее частотный вариант;
    - Случайных вариант;
    - Генерируем все возможные сочетания многозначных слов.

TurkLang – 7. 3 путь

Back-translation:

0. Сбор моноязычных корпусов и построения ЯМ:
    - Части параллельных корпусов
    - Моноязычные порталы
    - Существующие корпуса
1. Перевод промежуточными моделями (например, ru-tt):
    - Добавление полученных пар в обучающий корпус;
    - Обучение в обратном направлении (tt-ru).

Результаты
промежуточные

TurkLang – 7. Результаты. Киргизский

Направление перевода	Документов | Пар 2020	Документов | Пар 2021
https://sti.gov.kg/	934 | 4 178	284 | 980
http://www.kenesh.kg/	14 566 | 36 389	1 950 | 3 200
http://minjust.gov.kg/	3 578 | 17 092	520 | 77
http://novosti.kg/	90 595 | 36 517	19 500 | 7552
https://edu.gov.kg/ru/	2 414 | 21 140	
http://mineconom.gov.kg/ru	509 | 5 856	
http://med.kg/	684 | 1 370	
https://ru.sputnik.kg/news/	42 037 | 43 874	
JW300	- | 141 370	
https://saat.kg/	-	
http://kabar.kg/	-	
https://24.kg/kyrgyzcha/	-	

TurkLang – 7. Результаты. Башкирский

Направление перевода	Документов | Пар 2020	Документов | Пар 2021
bash.news	83 020 | 1 006	284 | 1 340
https://ufacity.info/	1 362 | 5 219	1 950 | 3 307
https://glavarb.ru/rus/	4 255 | 3 366	520 | 890
http://www.bashinform.ru/	> 500 тысяч | > 100 тысяч	- | 2 872
http://bashdram.ru/	2 383 | 2 122	| 124
https://house.bashkortostan.ru/	1 696 | 2 164	| 481
https://pravitelstvorb.ru/ru/	18 954 | 11 714	| 374
JW300	- | 47 658	-
Переведённая литерататура (6 книг)	- | 141 370	6 | 28 416

TurkLang – 7. Результаты. Татарский

Направление перевода	Документов | Пар 2020	Документов | Пар 2021
tatar-inform.tatar	877 327 | 708 665	42 767 | 57 810
https://tatarstan.ru/	167 206 | 232 075	946 | 7 198
https://kiziltan.rbsmi.ru/	41 954 | 12 166	9 477 | 2 930
JW300	- | 207 100	-
https://kzn.ru/	- | -	4 000 | 21 290

TurkLang – 7. Результаты. Узбекский

Направление перевода	Документов | Пар 2020	Документов | Пар 2021
https://kun.uz/	221 241 | 147 746	55 750 | 33 798
www.uzdaily.uz	52 176 | 52 117	7 508 | 4 961
https://www.gazeta.uz/en/	52 674 | 112 765	11 217 | 32 427
http://uza.uz/en/	42 174 | 113 705	35 712 | 58 303
http://xabar.uz/	3 258 | 14 532	3 468 | 8 092

TurkLang – 7. Результаты. Крымско-татарский

Направление перевода	Документов | Пар 2020	Документов | Пар 2021
Переведённая литература	7 | 1 109	14* (переразмечено) | 10 278
https://www.crimeantatars.club/	-	-

TurkLang – 7.

Ожидается пополнение базы благодаря проекту TIL:

1. Фильтрация по качеству.
2. Дедупликация.
3. Добавление языков для тестирования самых малоресурсных сценариев.

TurkLang – 7. Результаты «моноязычного» эксперимента

Направление перевода	BLEU ансамбля из 8 нейросетей	BLEU ансамбля из 4 нейросетей	BLEU M1; M2; M3; M4
Русско-башкирский	45.7	45.3	42.7; 43.8; 43.1; 42.6
Башкирско-русский	45.4	43.1	40.8; 40.2; 40.3; 40.0
Русско-киргизский	19.7	19.2	16.7; 17.6; 17.7; 18.4
Киргизско-русский	21.6	20.4	18.6; 18.0; 17.8; 18.5
Русско-чувашский	21.9	21.3	18.1; 18.3; 18.1; 18.2
Чувашско-русский	24.8	24.2	20.6; 20.9; 20.6; 20.6
Русско-казахский	48.2	49.0	45.6; 47.8; 43.9; 46.3
Казахско-русский	64.3	63.6	62.6; 62.3; 62.2; 62.2
Русско-крымско-татарский	13.5	13.8	12.3; 12.6; 12.5; 12.4
Крымско-татарско-русский	15.7	15.7	13.9; 13.5; 13.9; 14.3

TurkLang – 7. Результаты «многоязычных» экспериментов

Направление перевода	BLEU	ΔBLEU относительно базовых моделей
Русско-казахский	47.8	+0%
Казахско-русский	61.9	-1.1%
Русско-казахский (многоязычная)	48.4	+1.3%
Русско-татарский	33.6	+3.1%
Татарско-русский	36.4	+3.1%
Русско-татарский (многоязычная)	33.2	+1.8%
Русско-киргизский	22.2	**+20.7%**
Киргизско-русский	25.0	**+34.4%**
Русско-киргизский (многоязычная)	22.5	**+22.3**
Русско-узбекский	33.4	+9.9%
Узбекско-русский	35.5	**+11.3%**
Русско-узбекский (многоязычная)	31.1	+2.3%

TurkLang – 7. Результаты «многоязычных» экспериментов

Translation direction	BLEU	ΔBLEU относительно базовых моделей
Русско-башкирский	45.9	+4.8%
Башкирско-русский	47.3	**+15.9%**
Русско-башкирский (многоязычная)	47.3	+8.0%
Русско-чувашский	28.0	**+53%**
Чувашско-русский	30.4	**+45.4%**
Русско-чувашский (многоязычная)	25.8	**+41%**
Русско-крымско-татарский	22.7	**+80.2%**
Крымско-татарско-русский	24.4	**+70.6%**
Русско-крымско-татарский (многоязычная)	15.0	**+19%**

Спасибо!

Вопросы?

Контакт:

khusainov.aidar@gmail.com