МИНИСТЕРСТВО ОБРАЗОВАНИЯ РЕСПУБЛИКИ БЕЛАРУСЬ
БЕЛОРУССКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
ФАКУЛЬТЕТ ПРИКЛАДНОЙ МАТЕМАТИКИ И ИНФОРМАТИКИ
Кафедра дискретной математики и алгоритмики

ПРОКОПЕНКО Тимофей Александрович

МАШИННЫЙ ПЕРЕВОД ТЕКСТОВ МЕДИЦИНСКОГО И ЮРИДИЧЕСКОГО ТЕМАТИЧЕСКИХ ДОМЕНОВ

Магистерская диссертация

специальность 1-31 80 09 Прикладная математика и информатика

Научный руководитель
Гецевич Юрий Станиславович,
кандидат технических наук,
зав. лабораторией распознавания и синтеза речи ОИПИ НАН Беларуси

Допущена к защите
«___»_____ 2021 г.
Заведующий. кафедрой дискретной математики и алгоритмики
Котов Владимир Михайлович,
доктор физико-математических наук, профессор

Минск, 2021

ОГЛАВЛЕНИЕ

ПЕРЕЧЕНЬ УСЛОВНЫХ ОБОЗНАЧЕНИЙ

РБ – Республика Беларусь
МТД – медицинский тематический домен
НАН – Национальная академия наук
BLEU – bilingual evaluation understudy
NLP – natural language processing
BERT – bidirectional encoder representations from transformers
BPE – byte pair encoding
RNN – recurrent neural network

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Магистерская диссертация, 35 страниц, 16 рисунков, 4 таблицы, 15 источников.

МАШИННЫЙ ПЕРЕВОД, НЕЙРОННЫЕ СЕТИ, КОРПУСНАЯ ЛИНГВИСТИКА, ПЕРЕВОД НА БЕЛОРУССКИЙ ЯЗЫК, ЮРИДИЧЕСКИЙ ДОМЕН, МЕДИЦИНСКИЙ ДОМЕН

Объектом исследования магистерской диссертации являются задачи перевода текстов медицинской и юридической тематики с русского языка на белорусский и связанные с рассматриваемой задачей создание паралелльных корпусов, подготовительные этапи обработки текста и нейросетевые модели для перевода.

Целью работы изучение современных методов построения автоматических систем перевода, реализация собственной модели, оценка полученных результатов.

Методы исследования представляют собой анализ проблематики и существующих подходов, эксперимент, тестирование и сравнение.

Основные результаты исследования:

1. Проведен обзор теоретического материала, связанного с текущим развитием машинного перевода;
2. Созданы два новых параллельных корпуса медицинского юридического тематических доменов;
3. Разработан сервис с двумя различными интерфейсами, позволяющий переводить тексты различной тематики, а также проводить последующую адаптацию при появлении новых данных.

Областью применения разработанной модели являются прикладные задачи, связанные с переводом юридических и медицинских текстов с русского языка на белорусский.

АГУЛЬНАЯ ХАРАКТАРЫСТЫКА РАБОТЫ

Магістэрская дысертацыя, 35 старонак, 16 малюнкаў, 4 табліцы, 15 крыніц.

МАШЫННЫ ПЕРАКЛАД, НЕЙРОННЫЯ СЕТКІ, КОРПУСНАЯ ЛІНГВІСТЫКА, ПЕРАКЛАД НА БЕЛАРУСКУЮ МОВУ, ЮРЫДЫЧНЫ ДАМЕН, МЕДЫЦЫНСКІ ДАМЕН

Аб'ектам даследавання магістарскай дысертацыі з'яўляюцца задачы перакладу тэкстаў медыцынскай і юрыдычнай тэматыкі з рускай мовы на беларускую і звязаныя з разгляданай задачай стварэнне паралелльных карпусоў, падрыхтоўчыя этапы апрацоўкі тэксту і нейрасеткавыя мадэлі для перакладу.

Мэтай працы з'яўляецца вывучэнне сучасных метадаў пабудовы аўтаматычнага сістэм пераклада, рэалізацыя ўласнай мадэлі, ацэнка атрыманых вынікаў.

Метады даследавання ўяўляюць сабой аналіз праблематыкі і існуючых падыходаў, эксперымент, тэставанне і параўнанне.

Асноўныя вынікі даследавання:

1. Праведзены агляд тэарэтычнага матэрыялу, звязанага з бягучым развіццём вобласці машыннага перакладу;

2. Створаныя два новыя паралельныя корпусы медыцынскага і юрыдычнага тэматычных даменаў;

3. Распрацаваны сэрвіс з двума рознымі інтэрфейсамі, які дазваляе перакладаць тэксты рознай тэматыкі, а таксама праводзіць наступную адаптацыю пры з'яўленні новых даных.

Абласцямі ўжывання распрацаванай мадэлі з'яўляюцца прыкладныя задачы, звязаныя з перакладам юрыдычных і медыцынскіх тэкстаў з рускай мовы на беларускую.

SUMMARY

Master thesis, 35 pages, 16 figures, 4 tables, 15 sources.

MACHINE TRANSLATION, NEURAL NETWORKS, CORPUS LINGUISTICS, BELARUSIAN TRANSLATION, LEGAL DOMAIN, MEDICAL DOMAIN

The object of research of the master's thesis is the task of translating medical and legal texts from Russian into Belarusian and the creation of parallel corpora related to the task under consideration, preparatory stages of text processing and neural network models for translation.

The aim of the work is to study modern methods of building automatic translation systems, to implement the own model, to evaluate the obtained results.

Research methods are analysis of problems and existing approaches, experiment, testing and comparison.

Key results of the study:

1. A review of the theoretical material related to the current development of the field of machine translation was conducted;

2. Two new parallel corpuses of medical and legal thematic domains were created;

3. A service with two different interfaces developed, which allows you to translate texts of various topics, as well as carry out subsequent adaptation when new data appears.

Field of application of the developed model is applied problems related to the translation of legal and medical texts from Russian into Belarusian.

ВВЕДЕНИЕ

В Республике Беларусь документы преимущественно издаются на русском языке, что ограничивает их доступность для людей, которые не владеют или слабо владеют этим языком. С другой стороны, перевод медицинских и юридических текстов обязан быть точным, потому что от этого может зависеть жизнь человека. Следует также отметить, что на данный момент лингвисты затрачивают много времени для исправления однотипных ошибок в автоматическом переводе специализированных текстов. Для решения этих проблем был выбран путь создания собственной модели автоматизации перевода документов на другие языки, в частности, белорусский. Актуальность темы заключается в необходимости повышения уровня комфорта для всех категорий граждан, уникальность – в обработке словаря специфической профессиональной области. Целью данной работы будет сбор и обработка параллельных данных юридического и медицинского домена для русского и белорусского языка и реализация приложения для перевода.

В первой главе проведен обзор основных теоретических материалов на тему современного развития области машинного перевода, подробно описаны все этапы пред и постобработки текста для выбранной задачи, рассмотрена основная метрика для автоматического оценивания качества перевода.

Во второй главе приводится описание базового корпуса параллельных предложений, созданного ранее в рамках написания дипломной работы. Статья, описывающая процесс сбора данных, была представлена в марте 2021 года на V международной научно-практической конференции «Лингвистика, лингводидактика, лингвокультурология: актуальные вопросы и перспективы исследования» и ныне находится в процессе публикации. Также созданы два новых специализированных корпуса: юридический (на основе Кодексов Республики Беларусь) и медицинский (новостные данные, собранные с официальных сайтов учреждений здравоохранения).

В последней главе проведены эксперименты по тренировке собственной модели перевода, подготовлен сравнительный анализ полученных результатов, представлено два различных интерфейса для взаимодействия с моделью.

ГЛАВА 1
ОСОБЕННОСТИ ЗАДАЧИ

1.1 Обзор существующих алгоритмов машинного перевода

На данный момент выделяют несколько основных решений задачи машинного перевода. Перечислим их в хронологическом порядке и кратко опишем:

А) Машинный перевод на основе правил. Данный подход требует глубокого анализа на основе больших языковых моделей и правил и состоит из трех основных этапов.

1. Анализ исходного предложения. Включает обязательное определение частей речи, морфологический анализ (приведение слов к начальным формам), семантический анализ.

2. Трансфер. Перевод на основании словарей и правил, учитывая анализ первого шага.

3. Генерация. Дополнительное улучшение переведенного приложения, используя правила целевого языка (согласование в роде/падеже, порядок слов).

Очевидные минусы и сложности данного подхода:

1. Исключения;

2. Омонимы;

3. Необходимость обширной базы знаний об исходном и целевом языках.

Б) Статистический машинный перевод. Машинный перевод, основанный на применении статистических моделей, последовал за переводом на базе правил. Рассмотрим подвид статистического машинного перевода – перевод на базе фраз. Данный подход основан на трех основных частях:

1. Фразовая таблица, которая содержит переводы фраз и их вероятности для целевого языка;

2. Таблица перестановок слов, которая позволяет построить корректное предложение на целевом языке;

3. Языковая модель, которая может оценить вероятность существования каждой последовательности слов в целевом языке. Часто языковая модель просто проверяет онлайн-поиском и проверкой количества упоминаний.

Данный подвид перевода на нынешний момент редко используется в чистом виде, чаще в составе гибридных моделей. Статистический машинный перевод хорошо подходит для небольших предложений, но у него все еще слишком много проблем:

1. Согласование между фразами бывает плохим, так как генерация перевода происходит фактически с помощью конкатенации отдельных частей;

2. Не учитывается относительная важность слов, то есть более частотный перевод может быть выдан в качестве ответа без учета того, что, например, потеряна определяющая смысл частица *‘не’*;

3. Различные формы слов считаются разными словами.

В) Машинный перевод с помощью нейронных сетей. Стоит отметить, что с момента начала активного развития данного типа моделей (2015-2016 год), они практически сразу стали показывать наилучшие результаты на специализированных конференциях. Основная идея состоит в построении модели с использованием кодировщика и декодировщика. Предобработанные данные исходного языка кодируются в числовые векторы одинакового размера, а после кодировщик рекуррентно создает один итоговый вектор и подает его на вход декодировщика. Декодировщик, в свою очередь, генерирует перевод на целевой язык, учитывая вектор, полученный от кодировщика, и все предыдущие сгенерированные слова.

Базовой архитектурой для первых моделей были однонаправленные рекуррентные нейронные сети. Серьезный недостаток состоял в том, что контекст учитывался только для слов, стоящих перед переводимым словом. Позже идея была развита в виде двунаправленных RNN, которые учитывают и контекст после слова.

Построение переводчика данного типа прекрасно решает серьезные проблемы предшественников:

1. Различные формы слова будут закодированы похожим образом;

2. Учитывается контекст, а поэтому и относительная важность слов;

3. Тренировка требует лишь большого количества параллельных предложений, априорное знание об исходном и целевом языке не так важно.

Однако, можно заметить, что данный подход обладает большим минусом. Все представление входных данных кодируется в один вектор, таким образом в особенности для длинных предложений теряется часть информации и качество перевода падает. Эту проблему решило изобретение модели внимания (attention) в 2014 году [1]. Основое отличие состояло в генерации целевого предложения с использованием взвешенной комбинации векторов из кодировщика вместо использования единственного вектора фиксированного размера. На рисунке 1.1 представлена визуализация весов, на которой можно увидеть, какие именно слова исходного предложения наиболее важны для генерации слов на целевом языке.

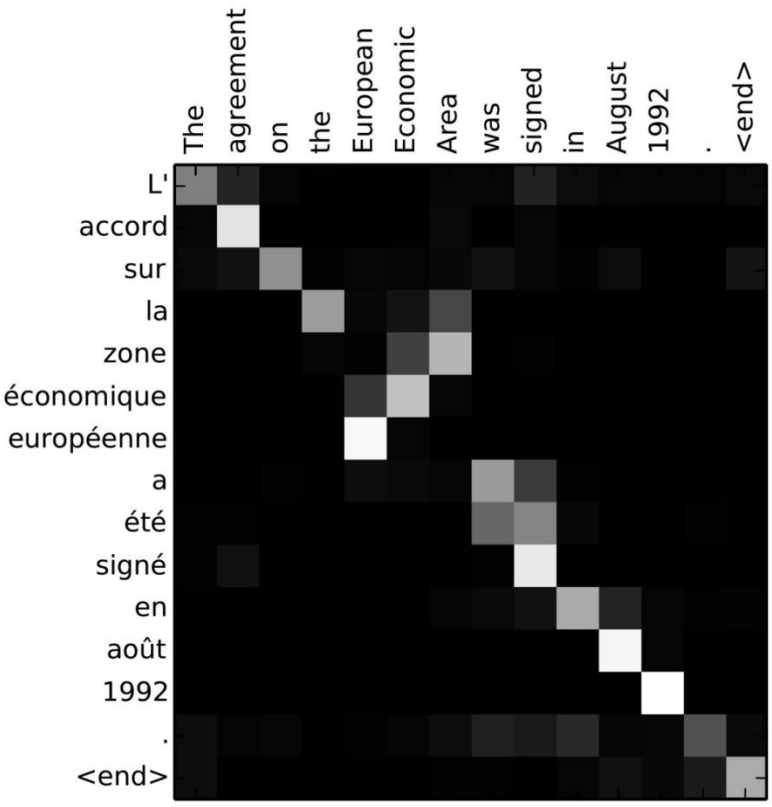

Рисунок 1.1 - Визуализация модели внимания

На рисунке 1.2 изображена архитектура Transformer. Основное новшество – это отказ от традиционных подходов (сверточные и рекуррентные сети) и изобретение механизма внутреннего внимания (self-attention). Данный слой помогает энкодеру посмотреть на другие слова во входящем предложении во время кодирования конкретного слова. Похожий механизм с небольшими изменениями используется и при передаче информации от кодировщика к декодировщику, а также в процессе декодирования. Из интересных моментов также стоит упомянуть применение множественного внимания (multi-head attention). Рассчитывается несколько представлений одних и тех же входных данных с помощью перемножения на разные матрицы, которые задают множественные «подпространства представлений». Благодаря данному подходу повышается способность модели фокусироваться на разных позициях в исходном предложении, а не всегда отдавать наибольший вес слову, кодируемому в данный момент.

Transformer улучшает качество перевода благодаря решению двух проблем. Во-первых, обычные рекуррентные модели даже с применением механизма внимания плохо справлялись с логической связью слов, которые находятся в предложении далеко друг от друга *‘The animal didn't cross the street because it was too tired’* у автоматического переводчика могут возникнуть проблемы с проблемы с определением того, к какому слову

относится местоимение ‘it’. Во-вторых, архитектура Transformer показывает очень высокой эффективность в условиях параллелизации, что позволяет значительно сократить время тренировки модели.

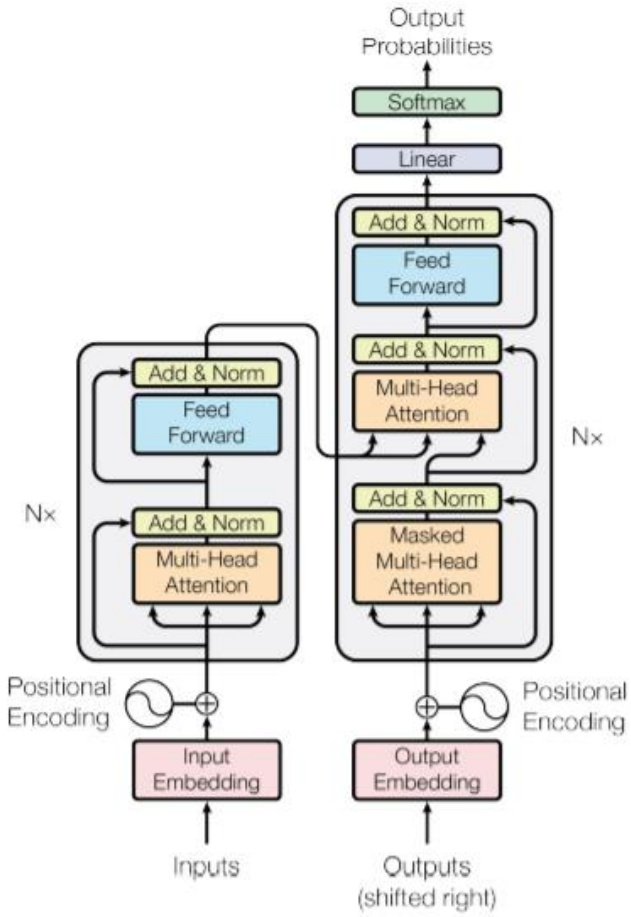

Рисунок 1.2 – Архитектура Transformer

Из последних исследований я хотел бы выделить два направления. Во-первых, стоит отметить успех моделей, которые обучаются для перевода не конкретной пары языков, а сразу для многих. До недавнего времени мультиязычные переводчики, в основном, использовали английский как промежуточный язык. Новейшие исследования Facebook [3] позволили собрать датасет, достаточный для того, чтобы обучить по-настоящему мультиязычную модель, которая не использует английский язык как промежуточный для перевода. Во-вторых, наилучшие результаты для многих задач NLP показывает BERT [4]. Эта модель получается довольно объемной и долго обучается, но она позволяет улучшить понимание языка гораздо лучше своих предшественников. Изначально BERT предназначался для предсказывания пропущенного слова и вероятности того, что одно предложение следует за другим. Теперь данную модель пробуют адаптировать и для задачи перевода [5]. Основная идея состоит в том, что слова обычный эмбеддинг слова из исходного языка комбинруется с

эбмеддингом, полученным из BERT, что позволяет получить более качественное векторное представление токена, чем раньше.

Заканчивая обзор типов машинного перевода, отметим, что некоторые современные переводчики также используют гибридный подход [6]. Модели гибридного перевода генерируют несколько вариантов с помощью различных типов переводчика (нейронные сети и статистический перевод), а потом с помощью специальной метрики выбирают наилучший. Как правильно, статистические модели лучше справляются с короткими предложениями без сложной лексики, а нейронные сети лучше улавливают контекст предложения и связь между словами в более сложных случаях.

1.2 Обзор основных автоматических систем перевода

Рассмотрим основные системы автоматического перевода на данный момент и оценим их качество для специфического случая русско-белорусского перевода медицинской лексики.

Будем сравнивать ‘Google Translate’, ‘Яндекс Переводчик’ и ‘Белазар’. Все перечисленные системы автоматического перевода довольно хорошо подходят для выбранной языковой пары (русский и белорусский языки), но основной минус в том, что исходный код данных систем закрыт. Мы никак не можем повлиять на ошибки данных моделей при переводе специфической лексики (например, МТД).

Для примера рассмотрим предложение *‘У пенсионеров часто наблюдается высокое АД.’* Из контекста ясно, что в данном случае *‘АД’* – это не противоположность рая, а сокращение для *‘артериального давления’*. На это указывает средний род прилагательного ‘высокое’ и упоминание пенсионеров. Посмотрим как данное предложение переведут существующие системы машинного перевода (эксперимент проводился 15.12.2019).

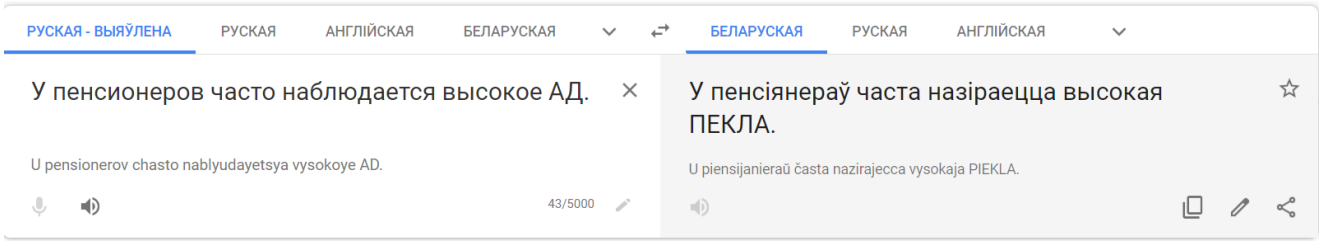

Рисунок 1.3 – Перевод ‘Google’ специфического предложения

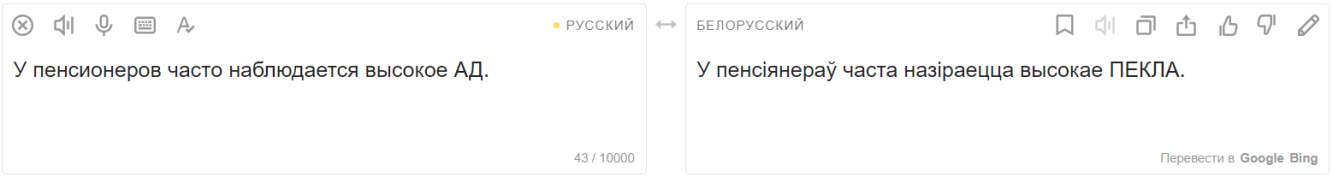

Рисунок 1.4 – Перевод ‘Яндекс’ специфического предложения

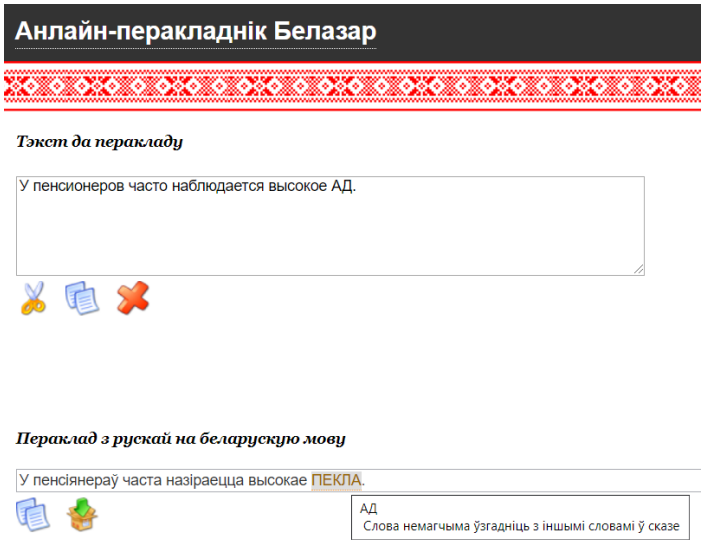

Рисунок 1.5 – Перевод ‘Белазар’ специфического предложения

Как мы видим, все три системы не смогли правильно перевести данное предложение. Стоит отметить, что ‘Белазар’ выдал предупреждение о несогласованности слов в предложении.

Несовершенство перевода именно для модели русский-белорусский показывает то, что данное предложение на английский и ‘Google’, и ‘Яндекс’ перевели корректно. Правда, ‘Яндекс’ по неизвестным причинам сохранил капитализацию слова *‘pressure’*:

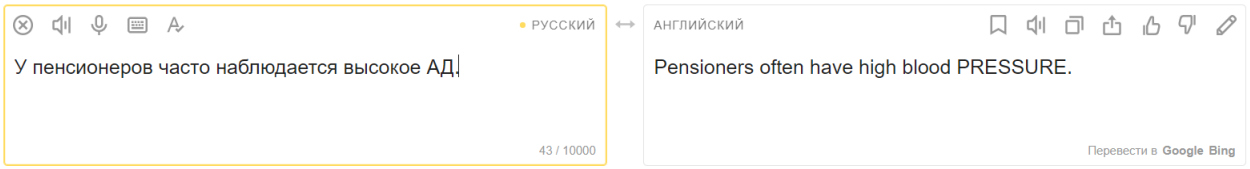

Рисунок 1.6 – Перевод ‘Яндекс’ специфического предложения на английский язык

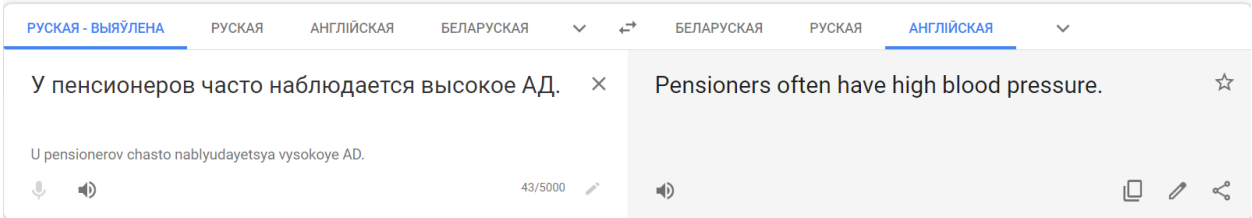

Рисунок 1.7 – Перевод ‘Google’ специфического предложения на английский язык

Таким образом, наличие собственной модели автоматического перевода, которую можно адаптировать при необходимости на специфические профессиональные области, является важной задачей.

1.3 Методы построения системы машинного перевода

Для построения собственной системы перевода были выбраны нейронные сети. Существует несколько готовых систем, которые позволяют обучить собственную модель, имея достаточное количество параллельных данных. Например, ‘OpenNMT’ [7], ‘Fairseq’ [8], ‘JoeyNMT’ [9].

Перед тем как переходить к переводу, необходимо обработать данные, а после представить итоговое предложение в виде, который ожидает пользователь. Основные этапы классической обработки: разбиение текста на предложения; отделение знаков пунктуации от слов; приведение больших букв к маленьким, где необходимо (трукейсинг). Если мы выбираем иную форму представление входного токена, чем слово, то также проводится сегментация.

Рассмотрим подробнее суть каждого из этапов и его важность.

1. Разбиение текста на предложения. Нейронная сеть учитывает контекст именно в рамках предложения, а не всего текста, поэтому данный этап ни в коем случае нельзя пропускать. Даже на очень длинных предложениях качество перевода падает, не говоря уже о переводе текста целиком. Данная задача имеет свои сложности. Например, точка может не являться знаком границы предложения, а обозначать сокращение или инициалы.

2. Отделение знаков пунктуации от слов. Важность данного шага кроется во внутренних представлениях кодировщика нейронной сети. Например, модель будет воспринимать слово *‘заяц’, ‘заяц.’, ‘заяц!’* как три разных слова, хотя и с похожим внутренним представлением. Очевидно, что для более качественной модели нужно представлять это слово абсолютно одинаково, а знаки пунктуации переводить отдельно.

3. Приведение регистра. Данный этап в какой-то мере схож с предыдущим, мы уменьшаем количество различных внутренних представлений модели путем приведения заглавных букв к строчным, где необходимо. Например, слово *‘Я’* и *‘я’* должны быть переведены одинаково, а токены *‘Варшава’* или *‘Купала’* к нижнему регистру приводить не нужно. Отдельно стоит отметить сложные случаи, когда имена собственные совпадают с именами нарицательными. К примеру, *‘Колас’* может быть как фамилией поэта, так и растением.

4. Сегментация. Для решения проблемы перевода слов, которых нет в словаре, построенном на исходных данных, был предложен алгоритм BPE [10]. Количество различных токенов, которое мы хотим видеть в итоговом разбиении, задается как параметр. После алгоритм изначально добавляет в словарь все возможные буквы, а далее в словарь добавляются самое частотное сочетаниие двух уже имеющихся токенов. Процесс

останавливается, когда набирается нужное количество токенов в словаре. Таким образом, самые популярные слова или, например, суффиксы будут представлены отдельным токеном, тогда как редкие слова будут переводиться практически на буквенном уровне.

После перевода все этапы повторяются в обратном порядке, чтобы итоговое предложение было в естественном для восприятия пользователем виде.

Для первых двух этапов часто используются специализированная библиотека ‘NLTK’ и ее подкомпоненты, а также библиотека ‘Sacremoses’.

Также интересные результаты показывает алгоритм ‘SentencePiece’ [11]. Это специализированный пакет для токенизации/детокенизации текста, который обучается на параллельных данных и строит собственную модель для разбиения предложения на токены. Данный подход также позволяет решить проблему ‘открытого словаря’, когда для корректного перевода нужно бесконечно расширять список известных модели слов. ‘SentencePiece’ сейчас используется во многих современных системах, хорошо интегрируется, например, с ‘JoeyNMT’. К важным достоинствам данного фреймоворка относится то, что 4 описанных выше этапа проводятся одной моделью. Таким образом, повышается качество детокенизации, а программисту не нужно комбинировать выходные данные нескольких различных программ.

В целях эксперимента обучены модели ‘SentencePiece’ для русского и белорусского языков на 60 тысячах параллельных предложений. Посмотрим на особенности разбиения предложения на токены.

```
['_чыста', 'я', '_ру', 'ская', '_', 'мова', '.']
чыстая руская мова.
['_ч', 'ы', 'ст', 'ая', '_ру', 'ская', '_м', 'ова', '.']
чыстая руская мова.
```

Рисунок 1.8 – Токенизация белорусского предложения

```
['_ч', 'и', 'сты', 'й', '_', 'рус', 'ск', 'и', 'й', '_я', 'з', 'ы', 'к', '.']
чистый русский язык.
['_чисты', 'й', '_русский', '_язык', '.']
чистый русский язык.
```

Рисунок 1.9 – Токенизация русского предложения

В обоих случаях первое разбиение порождает ‘белорусская’ модель, второе – ‘русская’ модель. Как видно, в зависимости от частоты встречаемости токенов в тренировочных данных порождается различное разбиение одного и того же предложения для моделей, обученных для разных языков.

1.4 Оценка качества модели

Оценка качества стала важной задачей с момента появления первых систем машинного перевода. Много исследований было посвящено поиску наиболее объективной метрики, но мы сейчас рассмотрим два основных подхода.

1. Экспертная оценка качества перевода. Этот подход до сих пор считается одним из лучших, хотя с ним же связана проблема субъективности в оценке качества. Машинный перевод почти всегда не идеален, поэтому требуются определенные знания и здравый смысл, чтобы оценить результат работы компьютерной программы. И там, где один специалист скажет, что перевод понятен и приемлем, для другого он будет «невозможен» с точки зрения стиля и грамматики. Один из способов борьбы с экспертной субъективностью – привлечение большого количества экспертов (или даже просто носителей языка).

2. BLEU [12] используется как основная автоматическая метрика. Очевидно, что применение экспертной оценки перевода замедляет проверку качества новых моделей, поэтому был разработан следующий алгоритм. BLEU оценивает качество перевода по шкале от 0 до 100 на основании сравнения человеческого перевода и машинного перевода и поиска общих фрагментов. Основная идея состоит в том, что чем больше совпадений, тем лучше перевод. Рассмотрим пример.

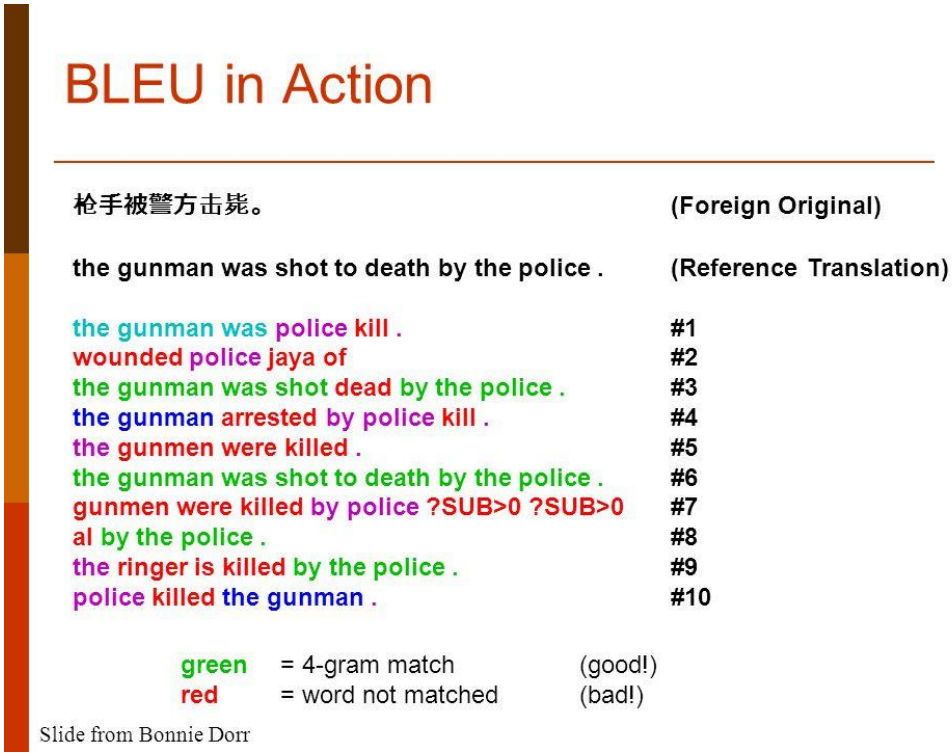

Рисунок 1.10 – Пример применения BLEU

На данном слайде из лекции Бонни Дорра мы видим, как именно выделяют n-граммные соответствия. И, очевидно, что 6 вариант машинного перевода является лучшим, так как совпадает с образцом, а 3 вариант также является качественным переводом из-за двух совпадающих 4-грамм.

Дополнительно учитывается «штраф за лаконичность». Например, если машинный перевод содержит только одно слово «the», то оно скорее всего будет встречаться во всех исходных вариантах, но такой перевод не является корректным. Для этого и вводится BP (Brevity penalty):

$$\mathrm{BP} = \min(1,\ \text{длина\_перевода}/\text{длина\_образца})$$

Если есть несколько идеальных образцов перевода, то для расчета BP выбираем длину образца, наиболее близкую к длине перевода.

Основная формула для расчета BLEU выглядит следующим образом:

$$\mathrm{BLEU} = \mathrm{BP} * \exp(\sum_{n=1}^{N} w_n \log p_n),$$

где w_n – это соответствующий вес,

p_n – точность по n – граммам.

У этой метрики есть свои недостатки:

1. Не отражает связность и согласованность перевода;
2. Не учитывает относительную важность слов (например служебные слова, синонимы, имена собственные, частица ‘не’);
3. Не всегда ее можно корректно интерпретировать;
4. Не учитывает возможность перестановки слов с сохранением корректности перевода.

Тем не менее, BLEU очень популярна среди специалистов, так как позволяет сравнивать разные системы или разные версии систем очень быстро и с приемлемым качеством.

1.5 Выводы

В первой главе получены следующие основные результаты:

– проведен теоретический обзор доступных алгоритмов машинного перевода, для дальнейней разработки выбраны нейронные сети;

– рассмотрены основные онлайн-переводчики для выбранной языковой пары на примере перевода конкретного предложения МТД;

– описан полный набор подзадач, необходимых для построения автоматического переводчика;

– приведено решение проблемы перевода слов, которых модель не видела во время обучения;

– описан основной алгоритм оценки качества модели машинного перевода.

ГЛАВА 2
ПОДГОТОВКА ДАННЫХ

2.1 Обзор базового корпуса

Для построения переводчика на основе нейронных сетей необходимо иметь достаточное количество параллельных предложений, причем чем больше данных, тем лучше обучается система. Для языков из разных групп для достижения приемлемого качества обычно берут около миллиона предложений, для похожих (таких как белорусский и русский языки) может быть достаточно четырехсот тысяч. Во время написания дипломной работы мной был собран новостной корпус с сайта БелаПАН [13]. Он состоит из четырехсот двадцати девяти тысяч параллельных предложений. Данные собирались с помощью веб-скрейпинга архивных новостей с 2007 по 2018 год. После тексты выравнивались в полуавтоматическом режиме, все спорные моменты решались вручную, таким образом, итоговое качество корпуса получилось высоким. В рамках подготовки магистерской диссертации работа по сбору корпуса была представлена на V международной научно-практической конференции «Лингвистика, лингводидактика, лингвокультурология: актуальные вопросы и перспективы исследования».

Существует также корпус, собранный в 2016 году на основе текстов Euroradio [14]. В виду похожести новостных материалов обучение базовой модели будет проводиться только на корпусе БелаПан, а дополнительные данные будут использованы на этапе тестирования для определения качества перевода.

2.2. Создание тематических корпусов

2.2.1 Обработка юридических текстов

Для доменной адаптации переводчиков также необходимы параллельные тексты определенной тематики. Объединенный институт проблем информатики НАН Беларуси реализует проект по обработке корпуса юридических текстов – Кодексов Республики Беларусь [15]. Данные представлены в виде параллельных файлов на русском и белорусском языках в формате doc и docx. Однако, для обучения нейронной сети лучше перевести тексты в формат txt и удостовериться в параллельности. Был применен следующий алгоритм обработки.

1. С помощью библиотеки ‘docx’ языка программирования Python считываем файлы в формат txt. Во время считывания разбиваем тексты по точке, точке с запятой, убираем заведомо некорректные предложения

(состоящие из одной точки или большого количества запятых, разделенных пробелами).

2. Проверяем структуру получившихся файлов вручную. Многие из них начинаются с перечисления ссылок на документы по изменению и дополнению кодекса. Эти данные не несут важной лингвистической информации, поэтому в итоговый файл начинаем записывать только после того, как встретим одно из слов из списка: ‘оглавление’, ‘содержание’, ‘раздел’, ‘глава’. Также нас не интересуют неуникальные предложения, поэтому мы храним множество уже обработанных данных в рамках одного текста. Дополнительно обрабатываются заголовки, так как они зачастую состоят из только заглавных букв.

После переходим к фазе полуавтоматического выравнивания. Так как исходные тексты переводились лингвистами и некоторые моменты можно было трактовать по-разному, кодексы на русском и белорусском оказались не всегда параллельными. Был применен скрипт, с помощью которого выравнивался базовый корпус. Ниже приведен алгоритм работы:

1. Запускаем скрипт. Он выводит индекс предложения, на котором начинается расхождение файлов, а также сами предложения и метаинформацию.

```
Coef: 0.29910714285714285
Border: 0.7841772
Index: 262
Законодательство, определяющее порядок административного процесса*

Працэсуальна-выканаўчы кодэкс Рэспублікі Беларусь аб адміністрацыйных правапаруш
эннях устанаўлівае парадак адміністрацыйнага працэсу, правы і абавязкі яго ўдзел
ьнікаў, а таксама парадак выканання адміністрацыйнага спагнання
```

Рисунок 2.1 Пример работы скрипта

2. Открываем соответствующие txt файлы и находим нужную строку. Копируем текст и смотрим, в чем причина расхождения с помощью поиска по подстроке в исходных docx файлах.

3. Вручную исправляем неточность и запускаем скрипт еще раз, пока файлы не станут полностью паралелльными.

Также все спорные моменты сохранялись в отдельный файл, чтобы после передать лингвистам для улучшения исходных данных. По итогу получился корпус из двенадцаць тысяч предложений, что вполне достаточно для проведения доменной адаптации.

2.2.2 Обработка текстов МТД

Медицинские тексты были представлены научным руководителем в формате txt (около пяти тысяч предложений). Это новости, собранные с специализированных сайтов (https://komzdrav-minsk.gov.by/, https://4gkb.by/). Данный небольшой корпус был также обработан, проведено выравнивание и чистка от повторяющихся предложений. Полученные доменные корпусы планируется использоваться в качестве тестовых и адаптационных данных.

2.3 Выводы

Для наглядности оформим размеры получившихся корпусов в таблицу и приведем примеры предложений.

Название корпуса	**Размер (тысяч предложений)**	**Примеры (русский)**	**Примеры (белорусский)**
Базовый корпус БелаПАН	429	Однако во многих странах переход на летнее время не осуществляется.	Аднак у многіх краінах пераход на летні час не ажыццяўляецца.
Юридический корпус	12	Привести решения Правительства Республики Беларусь в соответствие с настоящим Кодексом.	Прывесці рашэнні Урада Рэспублікі Беларусь у адпаведнасць з гэтым Кодэксам.
Медицинский корпус	4.5	Рядом - изотопная лаборатория, где выполняют радионуклидную диагностику...	Побач - ізатопная лабараторыя, дзе выконваюць радыенуклідную дыягностыку...

Таблица 2.1 – Обзор получившихся корпусов

Таким образом, во второй главе получены следующие основные результаты:

– произведен обзор базового корпуса, который будет использоваться для обучения модели;

– создан тематический корпус юридических текстов;

– создан тематический корпус медицинских текстов.

ГЛАВА 3
ПРОГРАММНАЯ РЕАЛИЗАЦИЯ СИСТЕМЫ

3.1 Апробация доступных алгоритмов и программных реализаций

Сначала было принято решение проверить концепцию перевода на небольшом корпусе из 65 тысяч предложений, причем их длина была ограничена 80 символами. Для построения собственной модели была выбрана библиотека ‘OpenNMT’. В качестве тренировочных предложений были использованы ранее собранные новостные данные с портала ‘БелаПан’.

Модель обучалась в течение двух суток. Итоговый результат выставлен в виде веб-сервиса. Пред- и постобработка данных проводилась с помощью ‘NLTK’.

Рассмотрим перевод некоторых предложений.

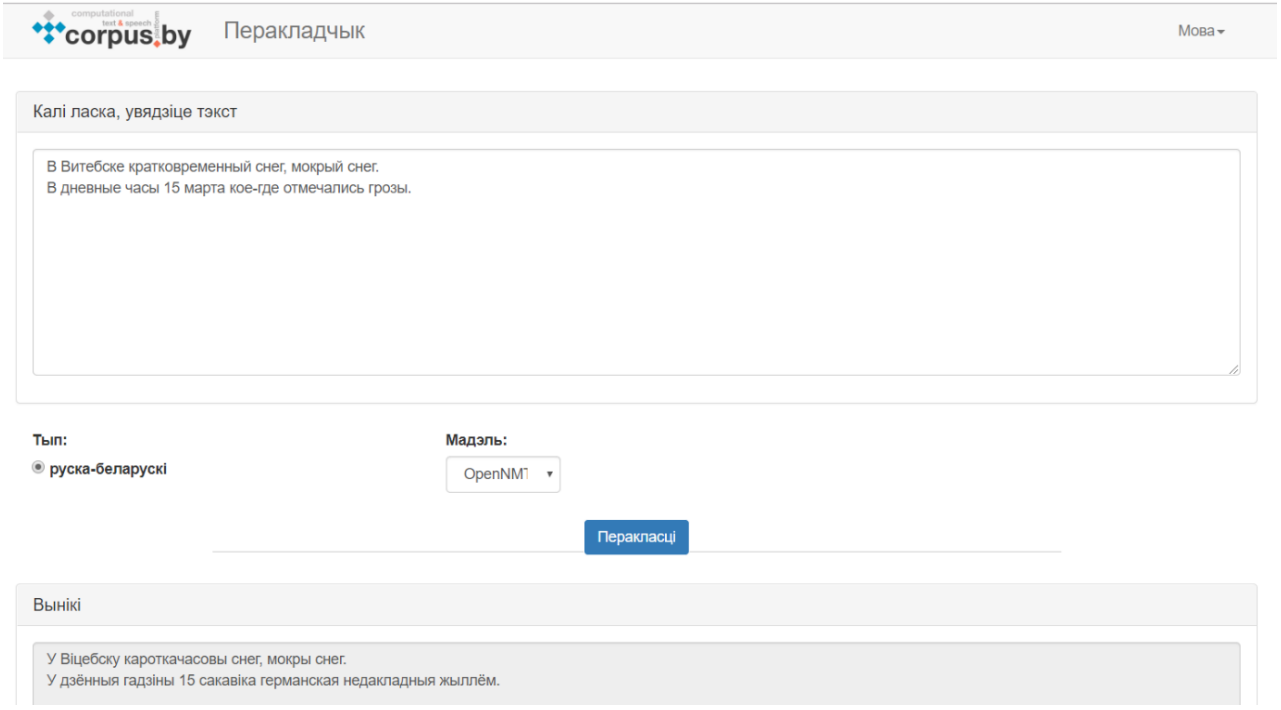

Рисунок 3.1 – Пример перевода с помощью собственной модели

Как видим, первое предложение переводится абсолютно корректно, а с переводом концовки второго уже возникают проблемы.

Проверим качество перевода на медицинских новостях.

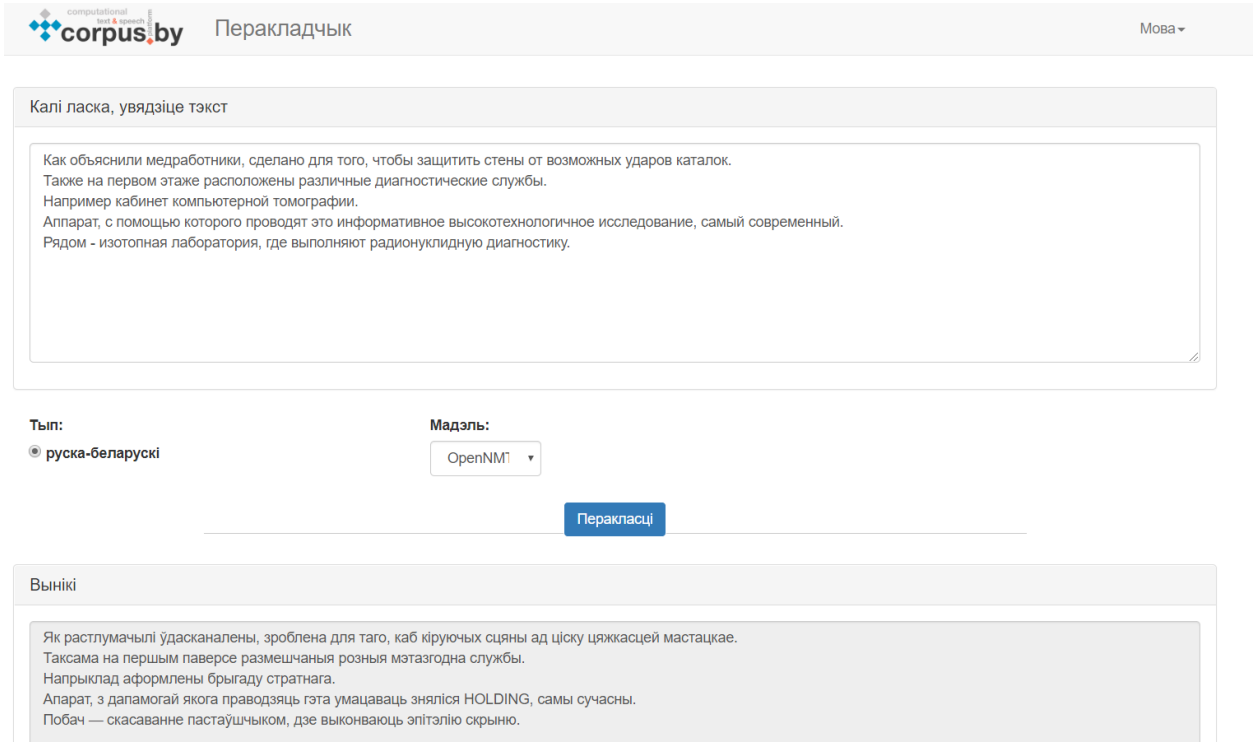

Рисунок 3.2 – Пример перевода МТД с помощью собственной модели

Медицинская лексика не входила в тренировочные данные, поэтому с этими словами возникают проблемы при переводе.

Таким образом, начальная версия сервиса выставлена на corpus.by, но еще предстоит значительная работа по улучшению качества перевода, исследованию других моделей и адаптации к тематическим доменам.

3.2 Определение текущего качества модели, выявление путей для улучшения

Начальная версия системы перевода, описанная в пункте 3.1 имела ряд недостатков:

1. Регистрозависимый перевод. Например, слова *‘привет’* и *‘Привет’* переводились по-разному, причем во втором случае перевод не был корректен;
2. Малый объем тренировочного корпуса (65 тысяч предложений) и, как следствие, слабая обобщающая способность модели;
3. Скорость перевода;
4. Низкие значения тестовых метрик (‘BLEU’).

Рассмотрим значения метрики ‘BLEU’ на тестовых выборках текущей реализации, чтобы можно было понять, насколько новая модель будет лучше:

1. Новостные данные МТД – 14.09;
2. Новостные данные ‘Euroradio’ (март 2016) – 19.91;

3. Тестовая выборка данных ‘БелаПАН’, на которых система обучалась – 28.07;

Стоит отметить, что для языков из разных групп значения данной тестовой метрики около 30 пунктов являются хорошим результатом. Для языков одной группы приемлемое значение гораздо выше – около 70 пунктов. Очевидно, что текущее решение нуждается в улучшении.

Новую модель будем разрабатывать на базе ‘JoeyNMT’. Данный модуль имеет подробную документацию и хорошо подходит для изучения особенностей моделей нейронного машинного перевода.

Также будет проведена тщательная пред- и постобработка входных данных, изменен уровень токенизации предложения на ‘BPE’ для более корректной обработки слов, не встречающихся в тренировочной выборке. Увеличим также объем тренировочной выборки до 430 тысяч предложений и максимальную длину предложения с 80 до 130 символов.

3.3 Разработка новой программной реализации

Разработка велась на языке программирования Python с помощью сервиса ‘Google Collaboratory’, были использованы библиотеки ‘sacremoses’ (обработка предложений), ‘subword-nmt’ (токенизация) и ‘JoeyNMT’ (перевод). Модель обучалась около двух суток. Проверим качество перевода тестовых предложений из пункта 3.1.

```
interactive_translate("В Витебске кратковременный снег, мокрый снег.")
interactive_translate("В дневные часы 15 марта кое-где отмечались грозы.")

У Віцебску кароткачасовы снег, мокры снег.
У дзённыя гадзіны 15 сакавіка я-дзе адзначаліся навальніцы.
```

Рисунок 3.3 – Пример перевода с помощью новой модели

```
[39] interactive_translate("Как объяснили медработники, сделано для того, чтобы защитить стены от возможных ударов каталок.")
     interactive_translate("Также на первом этаже расположены различные диагностические службы.")
     interactive_translate("Например кабинет компьютерной томографии.")
     interactive_translate("Аппарат, с помощью которого проводят это информативное высокотехнологичное исследование, самый современный.")
     interactive_translate("Рядом - изотопная лаборатория, где выполняют радионуклидную диагностику.")

     Як растлумачылі медработнікі, зроблена для таго, каб абараніць сцены ад магчымых удараў каталок.
     Таксама на першым паверсе размешчаны розныя дыягнастычныя службы.
     Напрыклад кабінет камп'ютарнай тамаграфіі.
```

Рисунок 3.4 – Пример перевода предложений МТД с помощью новой модели

Как видно, качество перевода значительно выросло, хотя остаются небольшие ошибки (*‘я-дзе’*). Особенно улучшение модели заметно на текстах МТД, которые ранее не переводились корректно.

Осталось проверить значения метрики ‘BLEU’ на ранее выбранных тестовых файлах.

Тестовые данные	**BLEU**
Новостные данные ‘БелаПАН’	75.309
Новостные данные ‘Euroradio’	69.861
Данные МТД	69.979
Юридические данные	73.634

Таблица 3.1 – Сравнение метрик качества на тестовых данных.

Дальнейшее направление исследования связано с разработкой полноценных демонстрационных программ.

Основным препятствием для презентации алгоритма на ‘Google Collaboratory’ является необходимость использования данных, которые на первой этапе разработки хранились на личном ‘Google Drive’. Эта проблема была решена с помощью облачного хранилища ‘Google Storage’. Необходимые файлы (модель, словарь, тренировочные тексты, файл конфигурации) были загружены на общедоступное внешнее хранилище. Благодаря приложению ‘gsutil’ они копируются в текущую сессию ‘Google Collaboratory’, демонстрационная программа, таким образом, становится автономной и готовой для презентации.

‘Google Cloud’ также предоставляет возможность построить свой переводчик с помощью сервиса ‘Google AutoML’. Он предназначен для обучения переводчика на базе системы ‘Google NMT’ (лежит в основе ‘Google Translate’). Данные трактуются как независимые пары, зависимости внутри текста не учитываются. Сервис делит корпус на тренировочную, валидационную и тестовую выборки, не учитывает дубликаты предложений. Предложения загружаются в формате tsv в специальный объект ‘DataSet’, после проводится тренировка.

Из плюсов ‘Google AutoML’ стоит отметить:

1. Простоту использования. Пользователь не занимается пост- и предобработкой текстов;
2. Улучшение качества базового ‘Google NMT’, что обеспечивает высокое качество итоговой программы-переводчика.

Минусы:

1. Закрытый исходный код;

2. Строгие ограничения на размеры тестовых/валидационных наборов данных (10 тысяч) и длину предложений (200 символов);

3. Ограниченных набор базовых пар для перевода.

К сожалению, данный сервис не применим к описываемой задаче, так как исследуемая языковая пара (русский и белорусский языки) недоступна в качестве базовой, а ручное расширение базовых пар не предусмотрено.

3.4 Доменная адаптация модели

Идея доменной адаптации состоит в том, чтобы дообучить базовую модель с помощью небольшого количества узкоспециализированных текстов. Шаги для адаптации следующие:

1. Разделить новые данные на тестовую и тренировочную выборки. Из тренировочная выборки позже автоматически будет создана валидационная.

2. Исправить файл тренировочной конфигурации: прописать путь к базовой модели (*'load_model'*), уменьшить размеры групп данных на каждом шаге тренировки (*'batch_size'*) и валидации (*'eval_size'*). Также необходимо выставить флаг *'reset_best_ckpt'*, чтобы оценка качества производилась уже на новых данных.

3. Далее можно запускать тренировочную программу *'full_joey_gdrive_legal_domain_adaptation.ipynb'*, выложенную в свободный доступ [11]. После окончания работы необходимо выложить полученную модель на 'Google Storage' для создания демонстрационной программы.

Результат метрики 'BLEU' новой модели на ранее выбранных тестовых файлах МТД вырос до 83.017.

Для наглядности полученные результаты были оформлены в виде таблицы сравнения перевода конкретного предложения.

Вид предложения	**Текст**
Русский язык (базовое)	Заболевание *относится* к группе аутоиммунных, то есть возникает, когда иммунная система, в норме *защищающая* организм от вредных агентов, например, бактерий и вирусов, ошибочно начинает атаковать здоровые клетки и *ткани*.

Таблица 3.2 Сравнение перевода двух моделей для данных МТД

Вид предложения	**Текст**
Белорусский язык (базовая модель без адаптации)	Захворванне *ставіцца* да групы аўтаіммунных, гэта значыць узнікае, калі імунная сістэма, у норме *наяўная* арганізм ад шкодных агентаў, напрыклад, бактэрый і вірусаў, памылкова пачынае атакаваць здаровыя клеткі і *тканіны*.
Белорусский язык (модель после адаптации)	Захворванне *адносіцца* да групы аўтаіммунных, гэта значыць узнікае, калі імунная сістэма, у норме *абараняе* арганізм ад шкодных агентаў, напрыклад, бактэрый і вірусаў, памылкова пачынае атакаваць здаровыя клеткі і *тканкі*.
Белорусский язык (экспертный перевод)	Захворванне *адносіцца* да групы аўтаіммунных, г.зн. узнікае, калі імунная сістэма, *якая* ў норме *абараняе* арганізм ад шкодных агентаў, напрыклад, бактэрый і вірусаў, памылкова пачынае атакаваць здаровыя клеткі і *тканкі*.

Таблица 3.2 Сравнение перевода двух моделей для данных МТД

Как видим, перевод модели без адаптации является довольно качественным, но есть определенные проблемы:

– неправильный перевод слова *‘относится’* в контексте (*‘ставіцца’* вместо корректного *‘адносіцца’*);

– ошибка на переводе причастного оборота, который в белорусском языке передается совершенно иной конструкцией (*‘наяўная’* вместо *‘якая абараняе’*);

– неправильное употребление слова *‘тканіна’*, которое в белорусском языке относится только к текстильным изделиям, а не биологическим тканям.

После адаптации специфические слова в контексте стали переводиться правильно, исчезла грубая ошибка при согласовании оборота (но употребленное в эспертном переводе местоимение ‘якая’ не появилось). Интересно также, что обе модели перевели оборот *‘то есть’* как *‘гэта значыць’*, хотя в образце употреблено сокращение *‘г. зн.’*.

Таким образом, концепция доменной адаптации была применена успешно. Перевод конкретных узкоспециализированных текстов улучшился, однако,

данный результат должен быть проверен и на других данных, так как предложений МТД было достаточно мало и могут иметь место искажения.

Далее был проведен эксперимент доменной адаптации нового корпуса – юридических данных. Алгоритм остался тем же самым, но результат метрики ‘BLEU’ для специфического корпуса поднялся до 87.266.

Вид предложения	**Текст**
Русский язык (базовое)	*Недропользователи* и иные лица, *обнаружившие* минералогические, палеонтологические и иные *уникальные* геологические материалы или *имеющие* сведения о них, сообщают *об этом* в Министерство природных ресурсов и охраны окружающей среды Республики Беларусь или его территориальные органы
Белорусский язык (базовая модель без адаптации)	*Нярокарыстальнікі* і іншыя асобы, *якія знайшлі* мінералагічныя, палеанталагічныя і іншыя *унікальныя* геалагічныя матэрыялы або *маюць* звесткі аб іх, паведамляюць *пра гэта* ў Міністэрства прыродных рэсурсаў і аховы навакольнага асяроддзя Рэспублікі Беларусь або яго тэрытарыяльныя органы
Белорусский язык (модель после адаптации)	*Нядрокарыстальнікі* і іншыя асобы, *якія выявілі* мінералагічныя, палеанталагічныя і іншыя *ўнікальныя* геалагічныя матэрыялы або *якія маюць* звесткі аб іх, паведамляюць *аб гэтым* у Міністэрства прыродных рэсурсаў і аховы навакольнага асяроддзя Рэспублікі Беларусь або яго тэрытарыяльныя органы
Белорусский язык (экспертный перевод)	*Нетракарыстальнікі* і іншыя асобы, *якія выявілі* мінералагічныя, палеанталагічныя і іншыя *ўнікальныя* геалагічныя матэрыялы або *якія маюць* звесткі аб іх, паведамляюць *аб гэтым* у Міністэрства прыродных рэсурсаў і аховы навакольнага асяроддзя Рэспублікі Беларусь або яго тэрытарыяльныя органы

Таблица 3.3 Сравнение перевода двух моделей для юридических данных

Разберем подробнее неточности перевода базовой модели на примере перевода конкректного предложения, приведенного в таблице 3.2:

– слово *‘недропользователи’*, очевидно, оказалось незнакомым для модели и поэтому был сгенерирован не самый корректный вариант (*‘нярокарыстальнікі’*);

– деепричастие *‘обнаружившие’* переводится корректно, однако, в данном контексте предпочтительнее использовать глагол *‘выявілі’*, а не *‘знайшлі’*;

– слово *‘унікальныя’* переведено без *‘ў’* после предыдущего слова, заканчивающего на гласную букву.

После доменной адаптации слово *‘недропользователи’* все еще осталось слишком сложным для правильного перевода, однако, остальные ошибки оказались исправлены. При этом несколько моментов (*‘имеющие’, ‘об этом’*) которые были переведены базовой моделью корректно, после адаптации были заменены синонимами, наиболее близкими к экпертному переводу. Данный факт можно объяснить тем, что юридические тексты обычно не допускают вариативности в формулировках в отличие от новостных материалов.

Таким образом, концепция доменной адаптации была успешно опробовано, качество модели на переводе специфических предложений значительно улучшилось по сравнению с базовым вариантом. Весь код, исходные данные и результаты выложены в открытый доступ [13].

3.5 Обновление сервиса на corpus.by

Для предоставления удобного доступа всем категориям пользователей было решено обновить начальную версию сервиса на сайте corpus.by. На базе демонстрационного сервиса, разработанного специалистами Объединенного института проблем информатики НАН Беларуси, приложение-переводчик было адаптировано с Ubuntu для Windows и получило удобный графический интерфейс. Доступны 3 режима перевода: базовый, медицинский, юридический. Также данный сервис поддерживает несколько языков локализации.

Пользователю предлагается ввести одно или несколько предложений на русском языке и выбрать тип модели. Далее производится разбиение введенного текста на отдельные предложения, токенизация каждого из них (отделение знаков пунктуации от слов), приведение слов к правильному регистру. Обработанные данные переводятся в формат BPE с помощью словарей, полученных на этапе тренировки базовой модели перевода. После происходит собственно перевод с помощью нейронной сети, детокенизация (удаление лишних пробелов между знаками пунктуации и словами) и обратное приведение регистра (например, возвращение капитализации для первого слова предложения).

К основным сложностям адаптации стоит отнести перенос всех этапов пред- и постобработки предложений в их оболочки для языка Python и установка некоторых библиотек (torch, torchtext) для Windows. Вид приложения представлен на рисунке 3.5.

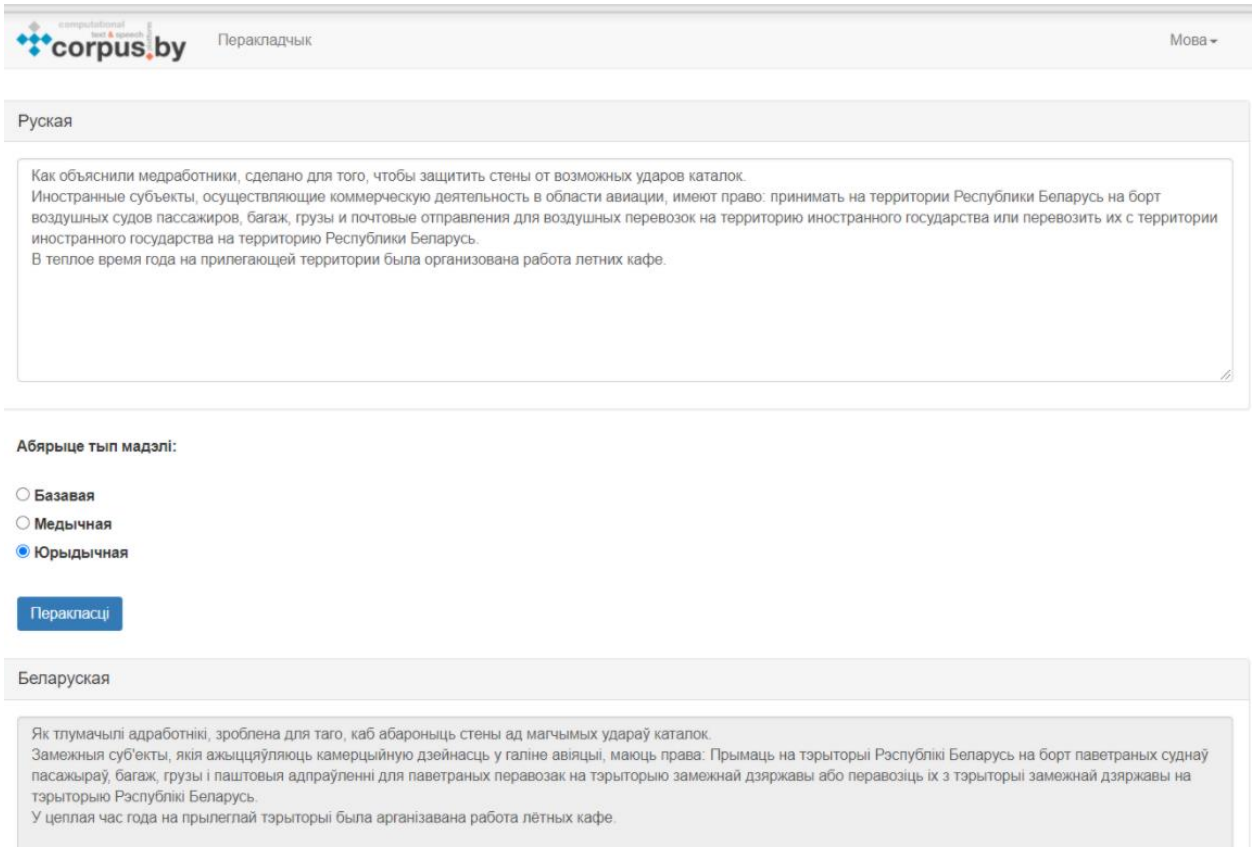

Рисунок 3.5 – Интерфейс сервиса ‘Translator’.

3.6 Выводы

В третьей главе получены следующие основные результаты:

– выполнен эксперимент с построением начальной версии переводчика, рассмотрены его слабые стороны, выявлены пути для улучшения;

– построена новая базовая модель со всеми этами пост- и предобработки входных данных;

– проведена доменная адаптация переводчика для юридических и медицинских данных, описан пошаговый алгоритм адаптации;

– определено качество перевода на каждом из этапов исследования;

– разработан демонстрационная программа на базе ‘Google Collaboratory’ и ‘Google Cloud’;

– сервис адаптирован для Windows и добавлен в набор программ, представленных на corpus.by.

ЗАКЛЮЧЕНИЕ

В процессе подготовки данной диссертации была решена задача по созданию нейронного машинного переводчика с русского на белорусский язык для текстов медицинского и юридических доменов. В частности, были созданы два новых корпуса специализированных текстов, натренирована базовая модель перевода. После была осуществлена доменная адаптация, представлен пошаговый алгоритм дообучения переводчика на новых данных. Сравнительный анализ результатов автоматической метрики BLEU показал хорошее качество получившихся моделей, что совпадает с экспертной оценкой лингвистов. Для взаимодействия с моделью были реализованы два различных интерфейса: микросервис, выставленный на corpus.by, а также демонстрационная программа на ‘Google Collaboratory’ с загрузкой данных из ‘Google Storage’.

Статья, описывающая процесс сбора и выравнивания данных, была представлена на V международной научно-практической конференции «Лингвистика, лингводидактика, лингвокультурология: актуальные вопросы и перспективы исследования» в марте 2021 года. Практические результаты данной работы внедрены в проект по усовершествованию автоматизированных систем обработки текстов юридической тематики, реализуемый Объединенным институтом проблем информатики НАН Беларуси.

Направления дальнейших исследований включают в себя ускорение скорости работы модели, сбор большего количества специализированных медицинских данных, а также применение других базовых фреймворков (Fairseq) и исследование новейших методов улучшения качества перевода (BERT NMT).

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. Bahdanau, D. Neural Machine Translation by Jointly Learning to Align and Translate / D. Bahdanau, K. Cho, Y. Bengio [Electronic resource] – 2014. Mode of access: https://arxiv.org/pdf/1409.0473.pdf. – Date of access : 03.02.2021.
2. Vaswani, A. Attention is all you need / A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin– 2017. – Mode of access: https://arxiv.org/pdf/1706.03762.pdf – Date of access: 15.02.2021.
3. Fan, A. Beyond English-Centric Multilingual Machine Translation / Fan A. [et al.] // Facebook AI [Electronic resource] – 2020. – Mode of access: https://ai.facebook.com/research/publications/beyond-english-centric-multilingual-machine-translation/ – Date of access: 16.02.2021.
4. Delvin, J. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding / J. Devlin, M.-W. Chang, K. Lee, K. Toutanova [Electronic resource] – 2018. – Mode of access: https://arxiv.org/abs/1810.04805/ – Date of access : 20.02.2021.
5. Zhu, J. Incorporating BERT into Neural Machine Translation / J. Zhu, Y. Xia, L. Wu, D. He, T. Qin, W. Zhou, H. Li, T.-Y. Liu [Electronic resource] – 2020. Mode of access: https://arxiv.org/abs/2002.06823. – Date of access : 25.02.2021.
6. Талбот, Д. Две модели лучше одной. Опыт Яндекс.Переводчика / Д. Талбот // habr.com [Электронный ресурс]. – 2018. – Режим доступа: https://habr.com/ru/company/yandex/blog/350002/ – Дата доступа: 02.03.2021.
7. OpenNMT-py: Open-Source Neural Machine Translation // github.com [Electronic resource] – 2017-2021. — Mode of access: https://github.com/OpenNMT/OpenNMT-py. – Date of access: 22.03.2020.
8. Fairseq(-py) // github.com [Electronic resource] — 2019-2021. — Mode of access: https://github.com/pytorch/fairseq. – Date of access: 25.04.2020.
9. Kreutzer, J. Joey NMT: A Minimalist NMT Toolkit for Novices // J. Kreutzer, J. Bastings, S. Riezler // [Electronic resource] – 2019. Mode of access: https://arxiv.org/pdf/1907.12484.pdf. – Date of access: 01.05.2020.
10. Sennrich, R. Neural Machine Translation of Rare Words with Subword Units / R. Sennrich, B. Haddow, A. Birch [Electronic resource] – 2016. – Mode of access: https://www.aclweb.org/anthology/P16-1162.pdf. – Date of access: 12.04.2020.
11. SentencePiece // github.com [Electronic resource] – 2020. Mode of access: https://github.com/google/sentencepiece/. – Date of access: 12.02.2020.
12. Papineni, K. Bleu: a Method for Automatic Evaluation of Machine Translation // K. Papineni, S. Roukos, T. Ward, W.-J. Zhu // [Electronic resource] – 2002. Mode of access: https://www.aclweb.org/anthology/P02-1040.pdf. – Date of access: 03.01.2020.

13. Translator [Электронный ресурс]. – 2019. Режим доступа : https://github.com/tsimafeip/Translator. – Дата доступа : 10.01.2021.
14. Euroradio [Электронны рэсурс]. – 2018. Рэжым доступу: https://euroradio.fm/cyaper-z-nulya-mozhna-lyogka-zrabic-autamatychny-perakladchyk-z-ruskay-na-belaruskuyu. – Дата доступу : 16.01.2019.
15. Удасканаленне працы аўтаматызаваных сістэм па тэкстах юрыдычнай тэматыкі // Аб'яднаны інстытут праблем інфарматыкі Нацыянальнай акадэміі навук Беларусі [Электронны рэсурс] – 2019-2021. Рэжым доступу: https://ssrlab.by/7804. – Дата доступу : 16.01.2021.