

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

САПАЧ Яна Петровна

**АЛГОРИТМЫ МАШИННОГО ПЕРЕВОДА ЮРИДИЧЕСКИХ
ТЕКСТОВ НА БЕЛОРУССКИЙ ЯЗЫК НА ОСНОВЕ
НЕЙРОННЫХ СЕТЕЙ**

Магистерская диссертация

специальность 1-31 81 09 «Алгоритмы и системы обработки больших
объёмов информации»

Научный руководитель
Юрий Станиславович Гецевич
доцент кафедры
биомедицинской информатики,
кандидат технических наук

Допущена к защите

«__» _____ 2020 г.

Зав. кафедрой дискретной математики и алгоритмики

_____ В.М. Котов

доктор физико-математических наук, профессор

Минск, 2020

ОГЛАВЛЕНИЕ

ПЕРЕЧЕНЬ УСЛОВНЫХ ОБОЗНАЧЕНИЙ	7
ВВЕДЕНИЕ	8
ГЛАВА 1. ОСОБЕННОСТИ ЗАДАЧИ.....	9
1.1 Основные понятия и определения	9
1.2 Особенности перевода текстов юридической тематики.....	10
1.3 Особенности перевода на белорусский язык	10
1.4 Обзор машинного перевода и его особенности	11
1.5 Обзор существующих решений и инструментов для машинного перевода текстов	12
1.6 Проблематика неизвестных слов	15
1.7 Проблематика форматирования текста	16
1.8 Выводы	16
ГЛАВА 2. ТЕОРИТИЧЕСКИЕ СВЕДЕНИЯ. ОБЗОР СУЩЕСТВУЮЩИХ МЕТОДОВ И ПОДХОДОВ К РЕШЕНИЮ ЗАДАЧИ	18
2.1 Обзор нейронные сети для машинного перевода	18
2.2 RNN.....	18
2.3 LSTM (Long Short Term Memory).....	19
2.4 CNN.....	20
2.5 Возможности решения проблемы неизвестных слов с помощью библиотеки SentencePiece.....	22
2.6 Возможности решения проблемы форматирования слов с помощью библиотеки stanza (StanfordNLP).....	22
2.7 Исследование LSTM с помощью библиотеки OpenNMT	23
2.8 Выводы	24
ГЛАВА 3. ПРОГРАМНОЕ ОБЕСПЕЧЕНИЕ	25
3.1. Описание исходных данных	25
3.2. Первая апробация	25
3.3. Использование корпуса «Еврорадио» и предобработка данных ..	26
3.4. Нормализация текста	28
3.5. Работа с редкими словами: добавление SentencePiece.....	33

3.6. Результаты экспериментов.....	34
3.7. Выводы	39
ЗАКЛЮЧЕНИЕ.....	42
СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ.....	43

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Магистерская диссертация, 44 с., 11 рис., 12 табл., 21 источник

НЕЙРОННЫЙ МАШИННЫЙ ПЕРЕВОД, НЕЙРОННАЯ СЕТЬ, ПАРАЛЛЕЛЬНЫЙ КОРПУС ТЕКСТОВ, ПЕРЕВОД НА БЕЛОРУССКИЙ ЯЗЫК, ЮРИДИЧЕСКИЙ ДОМЕН

Объектом исследования магистерской диссертации являются задачи перевода текстов юридической тематики, таких как кодексы Республики Беларусь, с русского языка на белорусский и связанные с рассматриваемой задачей подготовительные этапы обработки текста и нейросетевые модели для перевода.

Целью работы является построение нейросетевого переводчика текстов юридической тематики для белорусского языка.

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

Результаты исследования:

- определены и изучены основные трудности, связанные с нейросетевым переводом, переводом с русского на белорусский язык, переводом текстов юридической тематики;
- изучены наиболее распространённые подходы к машинному переводу текстов;
- изучены нейросетевые подходы к переводу текстов;
- подготовлены данные для тренировки моделей машинного обучения;
- сформирована схема работы с текстом, включающая себя стадии преобразования регистра, разделения предложений одного абзаца, разделения слов на более мелкие частицы;
- подготовлены модели для разделения слов на более мелкие частицы и для проведения обратного преобразования;
- подготовлена и обучена нейросетевая модель перевода;
- проведён сравнительный анализ экспериментов с разными подходами к обработке текстов.

Областью применения разработанной модели являются прикладные задачи, связанные с переводом юридических текстов, в первую очередь кодексов Республики Беларусь.

АГУЛЬНАЯ ХАРАКТАРЫСТЫКА РАБОТЫ

Магістарская дысертацыя, 44 с., 11 мал., 12 табл., 21 крыніца

НЕЙРОННЫ МАШЫНЫ ПЕРАКЛАД, НЕЙРОННАЯ СЕТКА, ПАРАЛЛЕЛЬНЫ КОРПУС ТЭКСТАЎ, ПЕРАКЛАД НА БЕЛАРУСКУЮ МОВУ, ЮРИДИЧНЫ ДАМЕН

Аб'ектам даследавання магістарскай дысертацыі з'яўляюцца задачы перакладу тэкстаў юрыдычнай тэматыцы, такіх як кодэксы Рэспублікі Беларусь, з рускай мовы на беларускую і звязаныя з разгляданай задачай падрыхтоўчыя этапы апрацоўкі тэксту і мадэлі для перакладу, заснаваныя на нейронных сетках.

Мэтай працы з'яўляецца пабудова нейрасеткавага перакладчыка тэкстаў юрыдычнай тэматыцы для беларускай мовы.

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

Вынікі даследавання:

- вызначаны і вывучаны асноўныя цяжкасці, звязаныя з нейрасеткавым перакладам, перакладам з рускай на беларускую мову, перакладам тэкстаў юрыдычнай тэматыцы;
- вывучаны найбольш распаўсюджаныя падыходы да машыннага перакладу тэкстаў;
- вывучаны нейрасеткавыя падыходы да пераклада тэкстаў;
- падрыхтаваны дадзеныя для трэніроўкі мадэляў машыннага навучання;
- сфарміравана схема працы з тэкстам, якая ўключае ў сябе стадыі прывядзення да каррэктнага рэгістру, падзялення сказаў аднаго абзацу, раздзялення слоў на дробныя часткі;
- падрыхтаваны мадэлі для раздзялення слоў на дробныя часткі і для правядзення зваротнага пераўтварэння;
- падрыхтавана і навучана нейрасеткавая мадэль перакладу;
- праведзена параўнальны аналіз эксперыментаў з рознымі падыходамі да апрацоўцы тэкстаў.

Абласцямі ўжывання распрацаванай мадэлі з'яўляюцца прыкладныя задачы, звязаныя з перакладам юрыдычных тэкстаў, у першую чаргу кодэксаў Рэспублікі Беларусь.

GENERAL CHARACTERISTICS OF THE WORK

Master thesis, 44 p., 11 figures, 12 tables, 21 sources

NEURAL MACHINE TRANSLATION, NEURAL NETWORK, PARALLEL CORPUS, TRANSLATION TO THE BELARUSIAN LANGUAGE, LEGAL DOMAIN

The object of the master's thesis is the problem of translating texts of the legal subject, such as codes of the Republic of Belarus, from Russian language into Belarusian and associated with the task text processing stages and neural network models.

The aim of the work is to build a neural network translator of legal texts for the Belarusian language.

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

The results of the study:

- the main difficulties associated with the neural network translation, translation from Russian to Belarusian, translation of legal texts were identified and studied;
- the most common approaches to machine translation of texts were studied;
- neural network approaches to translation of texts were studied;
- data for training machine learning models was prepared;
- a scheme for working with text was formed; it included the stages of register conversion, separation of sentences of one paragraph, separation of words into smaller pieces;
- models for dividing words into smaller pieces and for carrying out the inverse transformation were prepared;
- neural network translation model was prepared and trained;
- a comparative analysis of experiments with different approaches to word processing was carried out.

Field of application of the developed model is applied tasks related to the translation of legal texts, primarily codes of the Republic of Belarus.

ПЕРЕЧЕНЬ УСЛОВНЫХ ОБОЗНАЧЕНИЙ

ЕЯ – Естественный язык

РБ – Республика Беларусь

NN – нейронная сеть

CNN – свёрточная нейронная сеть

RNN – рекуррентная нейронная сеть

SMT – статистический машинный перевод

NMT – машинный перевод на основе нейронных сетей

ВВЕДЕНИЕ

Тема научной работы связана с применением новых достижений в области искусственного интеллекта для решения задачи машинного перевода юридических текстов. Ещё несколько лет назад основным способом машинного перевода был статистический анализ и использование словаря, однако большой скачок в развитии направления искусственных нейронных сетей привёл к новым разработкам для перевода текстов. Несмотря на то, что перевод публицистических текстов стал качественнее, в области узко профессионального перевода (юридический или медицинский домен) результаты не так убедительны. Помимо уникальных терминов, используемых в определённой профессиональной сфере, существуют так же слова и фразы, имеющие особый смысл для представителей узкого направления. Качество перевода становится критически важно.

В мульти языковых странах, таких, как Республика Беларусь, особенно остро ощущается нехватка ресурсов для быстрого перевода законодательных актов, кодексов и другой важной правовой информации на несколько языков. Вследствие этого понижается уровень доступности документов, падает уровень правовой образованности общества. Для решения проблемы перевода кодексов, актов и других текстов юридической тематики было решено исследовать возможности машинного перевода на основе нейронных сетей, а также изучить особенности белорусского и русского языков, проанализировать основные проблемы терминологии и её толкования в юридической сфере.

В первой главе сделан обзор литературы и существующих решений, сформулирована задача для выполнения в ходе магистерской работы.

В главе два проведён более детальный обзор некоторых существующих методов решения задачи, описаны программные возможности для решения сопутствующих проблем, описанных в первой главе, подготовлена база для программной реализации.

В главе три проведены практические эксперименты, описаны этапы развития итоговой модели, подготовлены значения метрики оценки и анализ результатов, выполнена реализация пользовательского интерфейса для взаимодействия с полученной моделью.

ГЛАВА 1

ОСОБЕННОСТИ ЗАДАЧИ

1.1 Основные понятия и определения

Естественный язык (ЕЯ) – язык, используемый людьми для общения.

Машинный перевод - процесс перевода текстов с одного естественного языка на другой с помощью специальной компьютерной программы.

Параллельные корпуса текста – часть текста, переведённый на два языка. Структура такого корпуса – отдельные фрагменты (обычно предложения), заканчивающиеся специальным символом, обозначающим конец фрагмента.

Искусственная нейронная сеть – программное воплощение биологической нейронной сети.

Нейронный машинный перевод (NMT) - подход к машинному переводу, в котором используется искусственная нейронная сеть.

Рекуррентная нейронная сеть (RNN) - вид нейронных сетей, где связи между элементами образуют направленную последовательность. Благодаря этому появляется возможность обрабатывать серии событий во времени или последовательные пространственные цепочки. В отличие от многослойных перцептронов, рекуррентные сети могут использовать свою внутреннюю память для обработки последовательностей произвольной длины [1].

Долгая краткосрочная память (LSTM) - разновидность архитектуры рекуррентных нейронных сетей, которая невосприимчива к длительности временных разрывов.

BLEU (оценка качества двуязычной замены – англ. *bilingual evaluation understudy*) – алгоритм оценки качества машинного перевода, в основе которого лежит оценка сходства машинного и человеческого перевода. За основу сравнения берётся количество вхождений n -грамм слов в машинном и человеческом переводе.

Для подсчёта BLEU по нескольким n -граммам, используется следующая формула:

$$BLEU = BP * \exp \left(\sum_{n=1}^N w_n * \log (p_n) \right), \quad (1.1)$$

где BP вычисляется по формуле:

$$BP = \begin{cases} 1, & \text{если } len(output_{mt-translated}) > len(output_{human-translated}) \\ \exp \left(1 - \frac{len(output_{mt-translated})}{len(output_{human-translated})} \right), & \text{в противном случае} \end{cases}, \quad (1.2)$$

а $w_n \in [0, 1], n \in [1, N]$ – веса различных n -грамм, $\sum_{n=1}^N w_n = 1$ [2].

1.2 Особенности перевода текстов юридической тематики

Перевод юридических текстов является одним из наиболее сложных видов письменного перевода. Связано это с тем, что существует необходимость в знании правовой области, а также в постоянном обновлении текущего профессионального словаря в связи с изменениями в законодательных актах. Отдельно стоит так же учесть особенности языков, которые используются для перевода, так как в исходном языке могут использоваться термины, которые нельзя описать, используя конструкции в языке, на который производится перевод [3].

Юридический текст требует особой точности в переводе терминов и формулировок, тут не допускается перевод приблизительный, по смыслу. Ведь даже мелкие неточности и едва заметные ошибки могут привести к неверному толкованию смысла документа и в итоге привести к неправильным действиям [4]. Неточность перевода терминологии или незнание особенностей правовой системы государства являются серьёзным препятствием для повышения количества юридических документов, переведённых одновременно на несколько языков.

1.3 Особенности перевода на белорусский язык

Перевод с белорусского языка и на белорусский представляет определенную сложность для носителей русского языка. Схожесть языков создает вероятность ошибок при переводе. То, что на русском будет считаться ошибочно записанным словом, на белорусском будет являться правильным. Особое внимание следует обращать на причастия:

- в белорусском языке нет страдательных причастий настоящего времени (нельзя, например, сказать «используемая литература»);
- для белорусского языка нехарактерно образование причастий с помощью суффиксов *-уч-* (*-юч-*), *-уц-* (*-юц-*), *-ем-* (*-ім-*), *-ач-* (*-яч-*); наличие таких причастий говорит о заимствованном характере слова;
- отсутствуют возвратные формы (с суффиксом *-ся*);
- не употребляются краткие страдательные причастия прошедшего времени в именительном падеже (например, «музей создан») [5].

У белорусского и русского языков очень много родственных межъязыковых омонимов. Белорусский язык очень похож на русский по своим лексическим и грамматическим единицам, которые похожи по форме, но значительно отличаются своим содержанием [6]. Для примера можно взять слово «трус», которое в русском языке обозначает боязливого человека, а в белорусском – кролика. Однако такая проблема как раз должна быть решена использованием машинного перевода.

1.4 Обзор машинного перевода и его особенности

Выделим несколько категорий машинного перевода:

- с постредактированием: исходный текст перерабатывается машиной, а человек-редактор исправляет результат;
- с предредактированием: человек приспособливает текст к обработке машиной (устраняет возможные неоднозначные прочтения, упрощает и размечает текст), после чего начинается программная обработка;
- с интерредактированием: человек вмешивается в работу системы перевода, разрешая трудные случаи;
- смешанные системы.

Для того, чтобы определить, какой именно тип машинного перевода подойдёт для решения поставленной задачи, проанализируем существующие условия. Приведём основные технологии перевода:

- машинный перевод на основе правил;
- статистический машинный перевод;
- машинный перевод на базе нейронных сетей;
- гибридный.

В первую очередь отметим, что для обучения нейронной сети необходимы большие объёмы тренировочных данных. Белорусский язык относится к так называемым языкам, с малым количеством подготовленных источников (*low-resource language*). Это значит, что ощущается нехватка информации и её придётся компенсировать или обрабатывать другими способами. Добавим сюда юридическую тематику и получим почти полное отсутствие благонадёжных текстов.

1.5 Обзор существующих решений и инструментов для машинного перевода текстов

На данный момент существует несколько решений для автоматического перевода текстов между русским и белорусским языками. Это Google Translate [7], Яндекс.Переводчик [8], Белазар [9]. Для тестирования переводчиков был взят отрывок из произведения «Людзі на балоце» Ивана Мележа. Результаты оценки качества перевода приведены на рисунках 1.1 – 1.3.

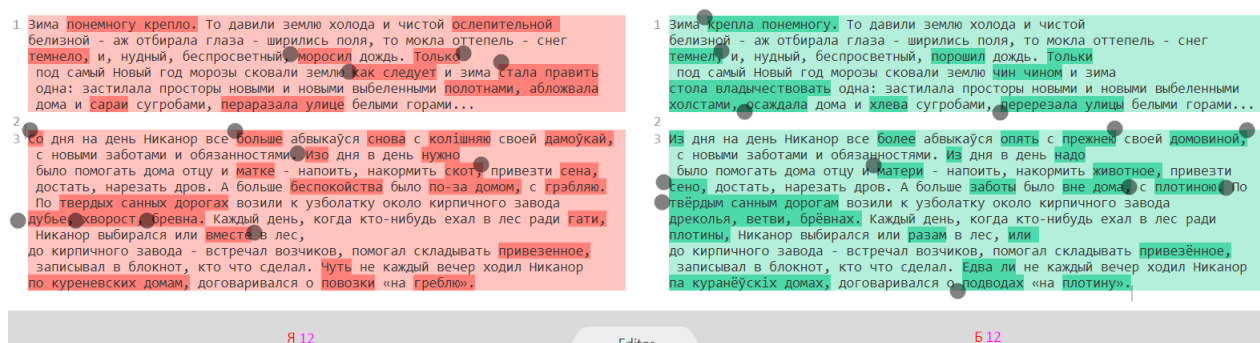


Рисунок 1.1 – Сравнение перевода Яндекса и Белазара. Чёрные точки указывают более подходящий вариант.

Видно, что Яндекс хуже знает уникальные для белорусского языка слова, однако Белазар тоже допускает промахи как с окончаниями, так и в принципе, с распознаванием слов. Оба перевода нуждаются в хорошей доработке.

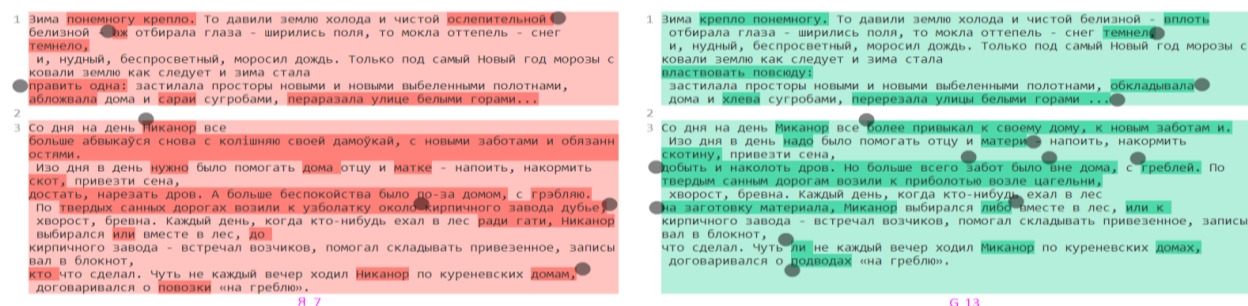


Рисунок 1.2 – Сравнение перевода Яндекса и Google. Чёрные точки указывают более подходящий вариант.

В паре Яндекс и Google явным лидером оказался второй участник. Некоторые части текста переведены столь литературно, что про машинный перевод можно даже не задуматься. Однако при сверке с оригиналом или переводом конкурента становится понятно, что для достижения благозвучности Google пренебрёг некоторыми словами. И если порой это приводило к блистательным переводам, то в некоторых случаях текст полностью терял смысл, так как было пропущено подлежащее.

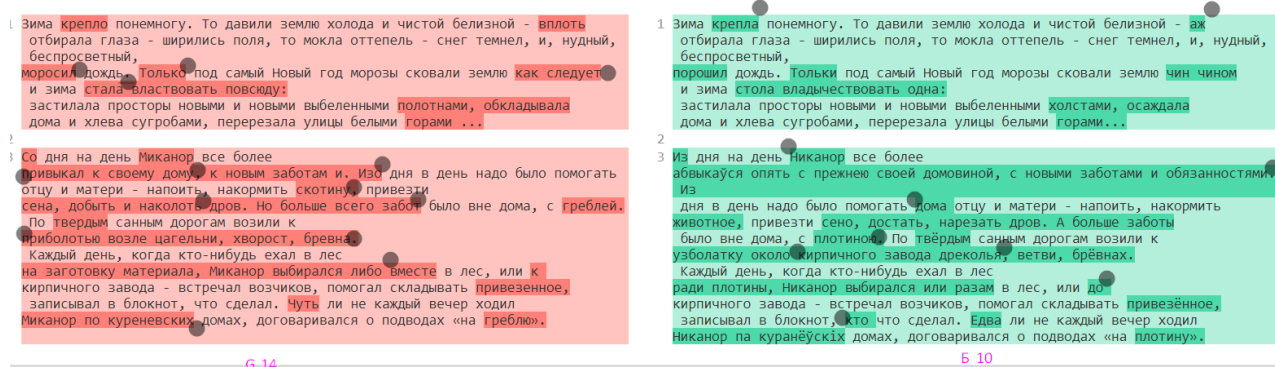


Рисунок 1.3 – Сравнение перевода Google и Белазара. Чёрные точки указывают более подходящий вариант.

В паре Google и Белазара перевод Google всё ещё первенствует, однако стоит заметить, что разница уже не кажется такой непоправимой.

Теперь проведём проверку на тексте юридической тематики. В данном случае был взят отрывок из кодекса РБ о внутреннем водном транспорте. В этот раз исследуем возможности перевести юридический текст с русского языка на белорусский. Результаты приведены на рисунках 1.4-6.

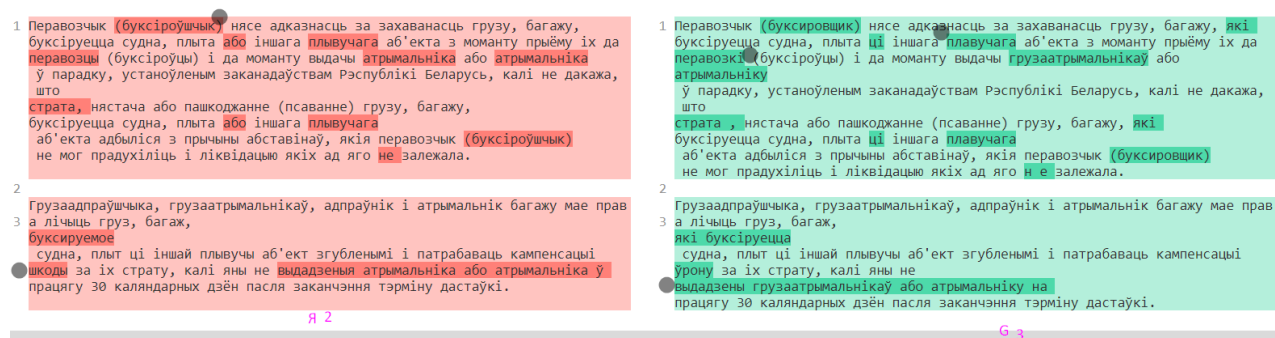


Рисунок 1.4 – Сравнение перевода Яндекса и Google. Чёрные точки указывают более подходящий вариант.

При просмотре перевода сразу бросается в глаза ошибки перевода узкоспециализированных слов (грузоперевозчик, буксировщик). Так же здесь видна особенность белорусского языка, описанная в пункте 2 главы 1, а именно: словосочетание «буксируемого судна» нельзя перевести дословно,

так как в белорусском языке такой формы не существует. Переводчики по-разному подошли к преодолению такого препятствия, однако ни у одного не получилось достичь приемлемого результата.

Если сравнить переводы Яндекс и Белазара, то здесь белорусская разработка показала себя лучше. С большего улучшения видны на наборе используемых слов для перевода (широте словарного запаса), а также на правильных окончаниях.

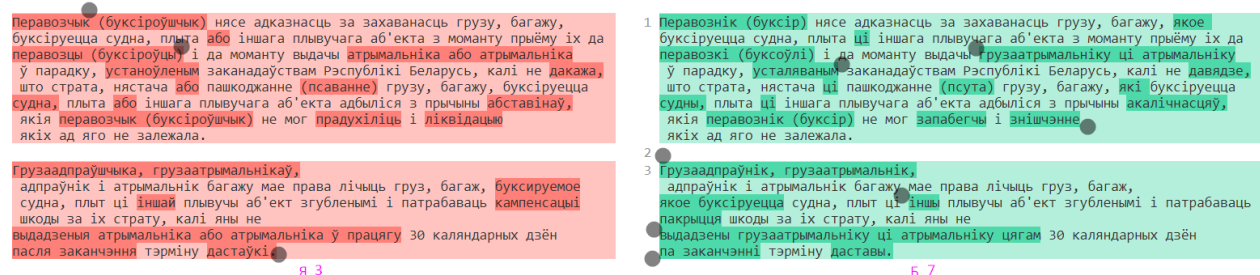


Рисунок 1.5 – Сравнение перевода Яндекс и Белазара. Чёрные точки указывают более подходящий вариант.

Google в этот раз не опустил важные части текста, что сделало перевод более точным, однако говорить о том, что какой-либо из кандидатов показал хороший уровень перевода, не приходится. Такой перевод не будет ясен и понятен читателю, что может привести к ошибочной трактовке кодекса в целом. Видно, что перевод нуждается в тщательной доработке.

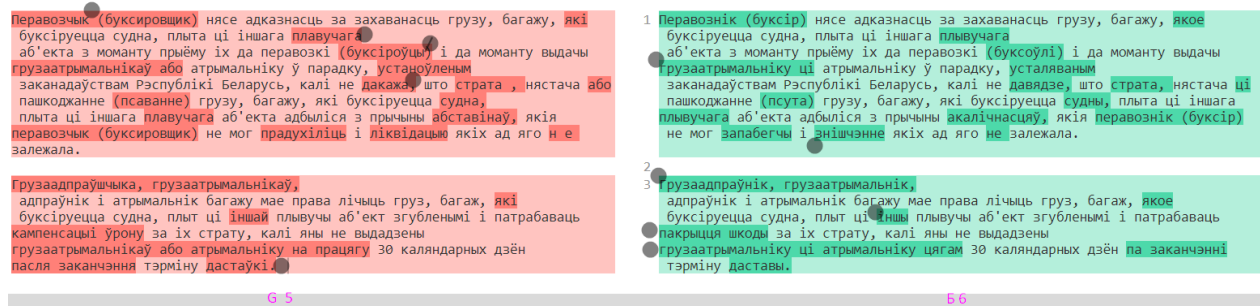


Рисунок 1.6 – Сравнение перевода Google и Белазара. Чёрные точки указывают более подходящий вариант.

Таким образом, на основе проведённых экспериментов, выявлена первая необходимость в расширении словаря специальных юридических терминов, который позволит производить точный перевод между языками. Следующими проблемами являются особенности самих языков и грамматических и лексических конструкций, а также расстановка правильных окончаний.

Источник [10] предоставляет доступ к словарю юридических терминов, с переводом на белорусский, русский и английский языки. Всего на момент

посещения в словаре находилось 27716 терминов. Не для всех терминов существуют переводы на все три языка. Некоторые, такие как «”зелёная” экономика», имеют только русский и белорусский вариант перевода.

Если обратиться к инструментам, существующим на данный момент для работы с машинным переводом, то будут выделены следующие разработки:

- NooJ – программа по разработке и использованию электронных словарей, которая так же даёт возможность производить маркировку текстовых корпусов и выводить текст в большое количество форматов [11];
- Translate – библиотека на языке Python для обучения моделей «sequence-to-sequence» [12]. Использует другую библиотеку Fairseq [13], которая так же направлена на тренировку моделей, связанных с обработкой текста, в том числе и его переводе;
- spaCy – библиотека для обработки текста;
- polyglot – ещё одна библиотека на языке Python с широким набором средств обработки от токенизации и транслитерации до морфологического анализа и распознавания частей речи.

1.6 Проблематика неизвестных слов

Отдельным вопросом стоит обработка переводчиком неизвестных(редких) слов. Для машинного перевода существует несколько подходов к неизвестным ранее словам:

- замена слова специальным тегом, одним для всех неизвестных слов / перенос слова в том виде, в котором оно встретилось в тексте-источнике. Очевидный минус такого решения – отсутствие попытки перевода;
- подготовка словарей с редкими словами и обновление их с помощью человека. Таким образом при последующей встрече слова перевод возможен, однако вариации слова не будут переведены, если их не оказалось в словаре. К тому же, требуется вмешательство человека;
- представление слов в виде групп букв, например «ночной» как «ноч», «н», «ой». Раздробление слова на более мелкие части позволяет сооружать возможный перевод и для неизвестных слов, так как исчезает само понятие «неизвестное слово» - все слова – это набор известных комбинаций букв алфавита. Такой подход хорошо сработает для вариации слова, например, когда

- существительное NN известно, а прилагательное встретилось впервые, или для составного слова, содержащего в себе уже известные (например, слово «огнеопасно»). Слабым местом такого подхода являются так называемые слова-исключения, модификация которых отличается от традиционной в языке;
- наиболее трудоёмкий в реализации – вариант с дополнительной NN, которая учится распознавать синонимы и, соответственно, при невозможности перевести слово напрямую, использовать для помощи существующую замену [14]. Подход так же можно организовать немного проще, через заранее подготовленные статические словари синонимов.

В следующей главе мы более подробно остановимся на возможностях предсказания незнакомого слова с помощью разделения его на более мелкие части.

1.7 Проблематика форматирования текста

Помимо невозможности объять все возможные конструкции языка ввиду его изменчивости, машинный перевод так же сталкивается с проблемой форматирования данных, которые к нему поступают. Предложение начинается с заглавной буквы, имена – тоже, а существуют ещё и аббревиатуры. Разница слов между «Ёж» и «ёж» не скажется на понимании текста человеком, но для NN это разные слова. При тренировке на таком материале качество будет ухудшаться, а при безоговорочном приведении к одному регистру будет потеряна часть аббревиатур, восстанавливать такой текст к исходному форматированию будет нелегко. Варианты возможных подходов к решению будут описаны в следующей главе, а в главе три будет проведён сравнительный анализ.

1.8 Выводы

В первой главе получены следующие основные результаты:

- даны определения основных терминов, используемых в работе;
- описаны уникальные особенности перевода юридических текстов;
- выделены критические моменты при переводе с русского на белорусский язык;
- описаны существующие подходы к решению задачи;

- проведён сравнительный анализ существующих крупных русско-белорусских переводчиков, выявлены их сильные и слабые стороны;
- описаны две крупные проблемы, которые необходимо учитывать при построении переводчика, представлены варианты их решения;
- приведён список существующих инструментов для программной разработки систем перевода.

ГЛАВА 2

ТЕОРИТИЧЕСКИЕ СВЕДЕНИЯ. ОБЗОР СУЩЕСТВУЮЩИХ МЕТОДОВ И ПОДХОДОВ К РЕШЕНИЮ ЗАДАЧИ

2.1 Обзор нейронные сети для машинного перевода

Впервые о NN для задач языкового перевода упомянули в 2014 в качестве альтернативы популярному статистическому переводу. Статистический перевод использовался крупными компаниями (такими, как Google, Microsoft, Facebook) в течение последних десятилетий, однако статья [15] полностью изменила принцип работы переводчиков. Недостатки SMT, такие как необходимость большого объёма обучающих данных, привлечение экспертов для каждой пары переводимых языков, сложность создания и поддержки системы в целом, привели к быстрому переходу от SMT к NMT.

2.2 RNN

Рекуррентные нейронные сети являются одним из видов NN. Их отличает от других NN использование предыдущих результатов для получения последующих, то есть «память». Это позволяет RNN работать с данными разной длины, то есть последовательностями данных, что является важной частью MT.

Общую схему RNN для NMT представлена на рисунке 2.1.

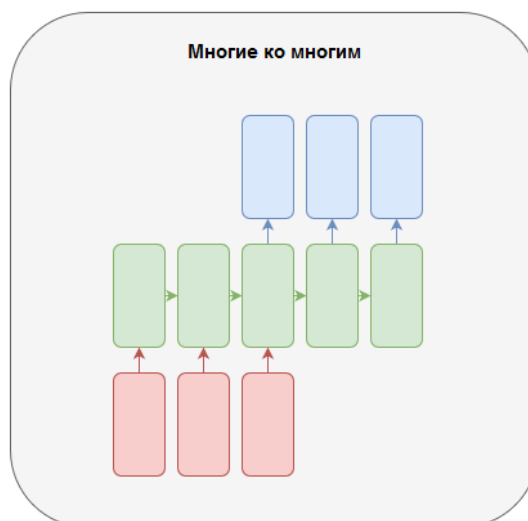


Рисунок 2.1 – Схема RNN для NMT. Красные прямоугольники – данные входной последовательности, зелёные – блоки RNN, синие – данные выходной последовательности.

На вход RNN подаётся вектор $\mathbf{x}$, сеть обрабатывает входные данные, используя при этом данные о своём внутреннем состоянии, и выдаёт вектор $\mathbf{y}$. Внутреннее состояние сети может изменяться при каждом шаге. В самом простом случае внутреннее состояние представляет собой вектор значений $\mathbf{h}$, первоначально инициализирующийся случайным образом (для простоты – нулями). Общая формула простой RNN:

$$\mathbf{h}_t = \tanh(\mathbf{W}_{hh}\mathbf{h}_{t-1} + \mathbf{W}_{xh}\mathbf{x}_t) \quad (2.1)$$

$$\mathbf{y}_t = \mathbf{W}_{hy}\mathbf{h}_t \quad (2.2)$$

Как и в случае вектора внутреннего состояния $\mathbf{h}$, обе матрицы заполняются случайными элементами. Одной из целей обучения RNN является подбор наиболее подходящих матриц.

Однако применение формул (2.1) и (2.2) в исходном виде не принесёт хороших результатов. В первую очередь, шагов в работе RNN некоторое количество, обычно достаточно большое (100 или 100000), на каждом шаге производится перемножение вектора с матрицей, что приведёт в конечном счёте либо к аннулированию, либо к ситуации очень больших чисел, что совершенно не поможет найти решение исходной задачи. Второй проблемой будет являться такая же ситуация с градиентом: и если переобученную модель можно будет подправить существующими методами, то при стремлении градиента к нулю вариантов выхода из ситуации нет. Таким образом, длина вектора входных данных будет существенно ограничена, что так же не приближает RNN к успеху. Решением проблем является улучшенная версия RNN.

2.3 LSTM (Long Short Term Memory)

Данный тип RNN обладает свойством запоминать информацию на долгое время. Общий вид LSTM представлен на рисунке 2.2.

В LSTM появляется дополнительный параметр – «ячейка» $\mathbf{C}$, на состояние которой влияют три шлюза (на рисунке 2.2 влияние шлюзов изображено стрелками). Такое влияние сети на состояние параметра $\mathbf{C}$ позволяет тому не уходить в крайности и оставаться относительно стабильным. С помощью первого шлюза (если рассматривать рисунок 2.2 слева направо) сеть регулирует, какое количество информации с предыдущих шагов

надо забыть, а оставшееся - пропустить далее. Это происходит путём передачи по шлюзу значения от нуля до единицы.

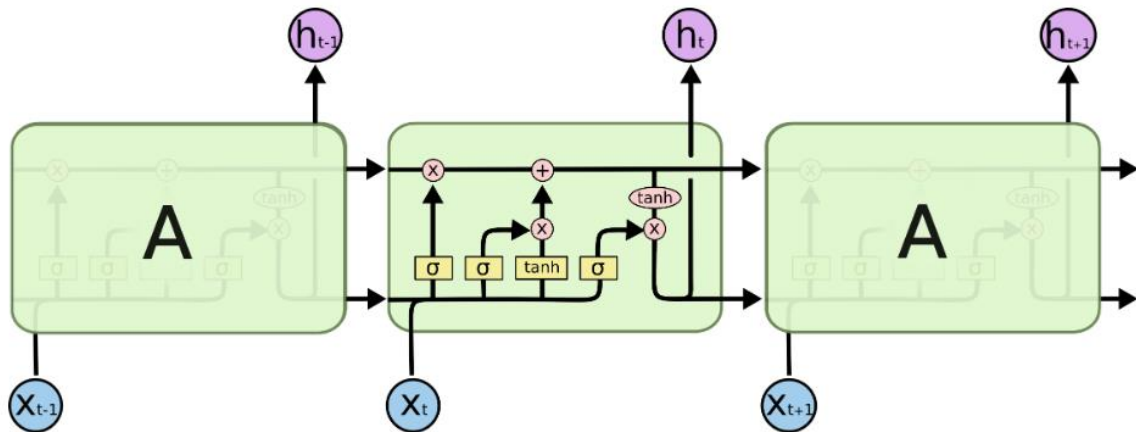


Рисунок 2.2 – Представление части LSTM сети.

Второй и третий шлюзы определяют, какую информацию необходимо положить в состояние C : второй определяет части состояния, подлежащие обновлению, третий – подготавливает вектор новых значений.

Таким образом, если обозначить значение в первом шлюзе через f_t , во втором - i_t и в третьем - $\tilde{C}_t$, новое состояние параметра C вычисляется по следующей формуле:

$$C_t = f_t C_{t-1} + i_t \tilde{C}_t \quad (2.3)$$

Четвёртый шлюз помогает высчитать новое значение h_t :

$$h_t = \tanh(C_t) * \sigma(W_o[h_{t-1}, x_t] + b_o) \quad (2.4)$$

2.4 CNN

Свёрточная нейронная сеть - специальная архитектура искусственных нейронных сетей, основной областью применения которой является задача распознавания изображений. Название архитектура сети получила из-за наличия операции свёртки, суть которой в том, что каждый фрагмент изображения умножается на матрицу (ядро) свёртки поэлементно, а результат суммируется и записывается в аналогичную позицию выходного изображения.

Потенциал CNN впервые был продемонстрирован исследовательским подразделением компании Facebook в 2017 году [16]. В отличие от RNN

свёрточные сети обрабатывают информацию одновременно, а не последовательно, что совершенно не похоже на поведение человека при чтении и переводе текстов. Для того, чтобы решить проблему, был предложен метод «многоуровневого внимания» (multi-hop attention).

Для каждого предложения, поступающего на вход CNN, рассчитывается представление каждого его слова w_j и его расположения в предложении l_j . На основании полученных представлений вычисляется слой пуллинга (подвыборки) z_i и состояние сети c_i :

$$e_j = w_j + l_j \quad (2.5)$$

$$z_i = \frac{1}{k} \sum_{t=-\lfloor \frac{k}{2} \rfloor}^{\lfloor \frac{k}{2} \rfloor} e_{j+t} \quad (2.6)$$

$$c_i = \sum_{j=1}^m a_{ij} e_j \quad (2.7)$$

На рисунке 2.3 приведена схема работы CNN.

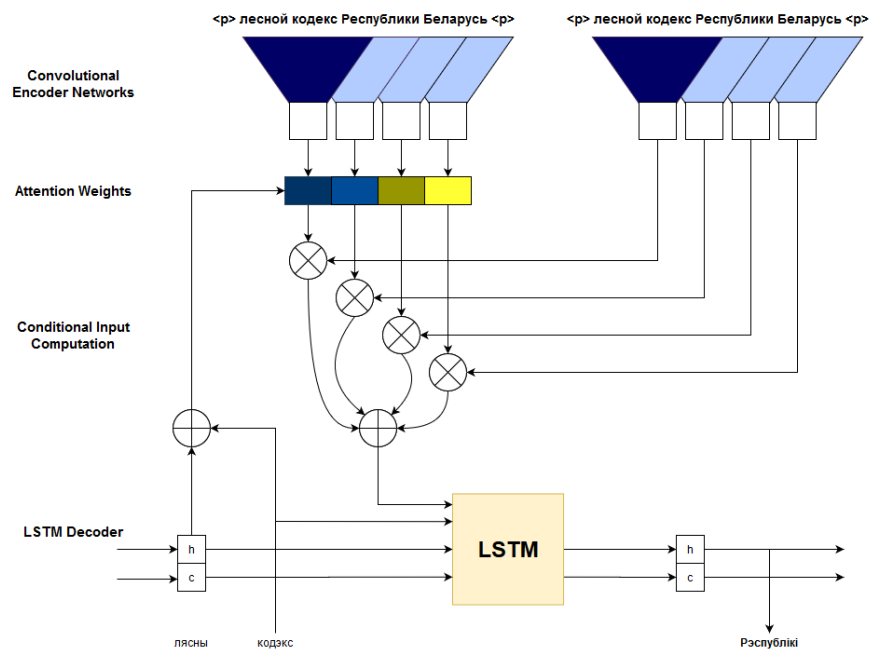


Рисунок 2.3 – Схема работы NN, использующей две CNN для кодирования предложения на ЕЯ и LSTM для декодирования на другой ЕЯ.

CNN для NMT представляет собой две свёрточные сети: первая сеть используется для получения коэффициентов весов представления отдельного слова, вторая – для генерации состояния системы.

Технология многоуровневого внимания позволяет CNN обращаться к различным частям предложения(текста), чтобы понять их смысл во время перевода. Такое возможно благодаря параллелизму CNN.

2.5 Возможности решения проблемы неизвестных слов с помощью библиотеки SentencePiece

Как уже отмечалось в первой главе, одной из острых проблем машинного перевода является перевод редких слов, которые NN не встречала при обучении. В данной работе мы остановимся на исследовании возможностей дробления слова на более мелкие частицы и влияние такого подхода на результат.

Существует два основных подхода к делению слов: BPE (англ. «Byte Pair Encoding» - кодирование парой байтов) и Unigram (униграмма), оба основываются на последовательном объединении более мелких частиц слова в более крупные. Алгоритмы завершают работу, когда достигают ограничения по количеству таких частиц слов. Предпочтение отдаётся тем частицам, которые чаще других встречаются в исходных данных, представляющих собой крупные корпуса текстов. Таким образом, для обучения алгоритма дробления для разных языков не нужно искать параллельные корпуса текстов, что даёт ему преимущество.

Рассмотрим реализацию обоих алгоритмов в библиотеке SentencePiece [17]. Конкретно, взглянем на сравнительную таблицу двух алгоритмов, приведённых в источнике [18]. Для разных языковых пар, приведённых в таблице, оба алгоритма имеют минимальную разницу, ощутимого перевеса не чувствуется. Мы остановимся на алгоритме Unigram. Важным параметром алгоритма является размер генерируемого словаря. В главе три мы рассмотрим вариации размера и оценим результаты.

2.6 Возможности решения проблемы форматирования слов с помощью библиотеки stanza (StanfordNLP)

Далее, в главе три, мы реализуем несколько подходов к форматированию текста, сейчас подробнее остановимся на определении корректного регистра. Для этого пригодится информация о том, к какой части речи принадлежит слово. В ходе исследования была найдена библиотека StanfordNLP, которая в

последствии преобразовалась в библиотеку stanza. В ней подготовлены алгоритмы определения части речи как для русского, так и для белорусского языка. Благодаря определению части речи слова, мы можем скорректировать его текущий регистр. Однако проблему с собственными существительными, совпадающими с нарицательными, решить не удалось, что говорит о несовершенстве подхода.

2.7 Исследование LSTM с помощью библиотеки OpenNMT

Для апробации нейронных сетей в переводе юридических текстов с русского языка на белорусский используем LSTM сеть разработки OpenNMT [19]. Архитектура представлена на рисунке 2.4.

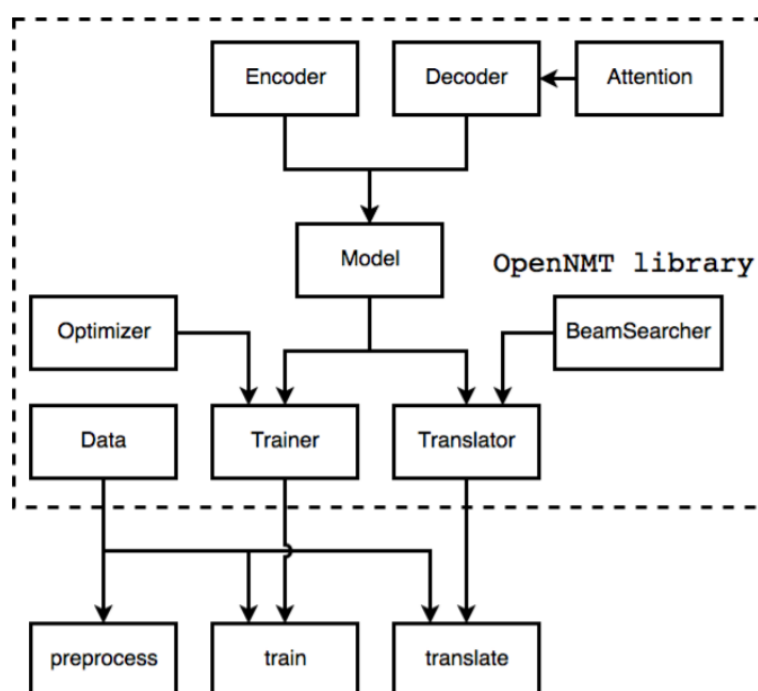


Рисунок 2.4 – Архитектура OpenNMT [20].

Сеть имеет два слоя с вектором состояния, насчитывающим 500 элементов. Для работы сети необходимо четыре файла: два для обучения и два для проверки корректности. Помимо этого, было создано два файла для тестирования результата обучения, на которых подсчитывалась метрика BLEU. Обучение моделей проводилось на разных платформах, время обучения, без учёта предварительной обработки данных, занимало от часа до девяти, в зависимости от количества шагов обучения и мощности предоставленного оборудования.

2.8 Выводы

Во второй главе получены следующие основные результаты:

- произведён обзор искусственных нейронных сетей, использующихся для перевода текстов;
- описан подход к решению проблемы неизвестных слов с помощью библиотеки SentencePiece;
- описан подход к решению проблемы корректного регистра слов с помощью библиотеки stanza;
- описана архитектура библиотеки для работы с нейронными сетями OpenNMT;
- определены пути реализации и перечень необходимых данных и экспериментов.

ГЛАВА 3

ПРОГРАМНОЕ ОБЕСПЕЧЕНИЕ

3.1. Описание исходных данных

Для того, чтобы данные подходили для тренировки нашей NN, необходимо выполнение двух условий:

- данные должны представлять собой параллельный корпус текстов на русском и белорусском языках;
- данных должно быть много.

Как уже отмечалось в первой главе, белорусский язык является малоресурсным, а, учитывая юридическую направленность перевода, круг данных сужается ещё больше. На протяжении всего периода исследования данные пополнялись. Поэтому в данной главе будут приведены результаты NN как на очень малом количестве данных, так и на более крупном.

В первую очередь, были подготовлены тексты на основе кодексов Республики Беларусь, для которых существовали подготовленные переводы: «Лесной кодекс», «Кодекс о земле», «Воздушный кодекс». Каждый кодекс содержит порядка 1000 строк, в каждой строке – одно или несколько предложений. В последствии исходные данные были улучшены и приведены к виду «одна строка – одно предложение».

Следующим этапом были добавлены параллельный корпус текстов «Еврорадио» [21], который содержит записи радио за несколько лет, с уже подготовленным форматом «одна строка – одно предложение». Исходные данные в корпусе разделены на 12 частей. В зависимости от этапа исследования, размер разделения на тренировочную и проверочную выборки, будет меняться. Всего в корпусе порядка 135000 предложений на каждом языке. Несмотря на общезыковую направленность корпуса, такой прирост в данных для обучения сказался очень существенно на качестве результатов.

3.2. Первая апробация

Обработка данных претерпела наибольшее количество изменений.

На ранних этапах она отсутствовала, что приводило к переводу, продемонстрированному в таблице 3.1.

Язык	Текст
<i>русский</i>	В связи с принятием настоящего Кодекса признать утратившими силу Кодекс о браке и семье Республики Беларусь, утвержденный Законом Республики Беларусь от 13 июня 1969 года (СЗ БССР, 1969 г., № 17, ст. 278), а также все акты законодательства, которыми внесены изменения и дополнения в Кодекс о браке и семье Республики Беларусь в период с 13 июня 1969 года до момента вступления в силу настоящего Кодекса.
<i>белорусский (NN)</i>	ДЗЯРЖАЎНАЯ 78. аб'екты за лесакарыстанне ЛЯСНЫХ РЭСURCAЎ ЛЯСНЫХ РЭСURCAЎ павінна змяшчаць: павінна РЭСURCAЎ і абарону дрэў ўчастка ляснага фонду павінна змяшчаць: павінна РЭСURCAЎ павінна змяшчаць: павінна змяшчаць: павінна змяшчаць: павінна РЭСURCAЎ павінна ляснага ляснага фонду; павінна змяшчаць: павінна РЭСURCAЎ павінна РЭСURCAЎ павінна РЭСURCAЎ павінна змяшчаць: павінна змяшчаць: павінна ляснага фонду; павінна змяшчаць: павінна ДРУГАРАДНЫХ ЛЯСНЫХ РЭСURCAЎ павінна РЭСURCAЎ ЛЯСНЫХ высечак ЛЯСНЫХ высечак РЭСПУБЛІКІ павінна РЭСURCAЎ
<i>русский</i>	Президент Республики Беларусь
<i>белорусский (NN)</i>	Прэзідэнт Рэспублікі Беларусь
<i>русский</i>	А.Лукашенко
<i>белорусский (NN)</i>	А.Лукашэнка

Таблица 3.1 – Результаты перевода с помощью обученной LSTM сети, обученной на одном из кодексов Республики Беларусь. Данные не проходили предварительную обработку.

Как видно из таблицы 3.1, малый размер тренировочных данных, уменьшенный многообразием регистров и отсутствием обработки неизвестных слов, привёл к слабым результатам: максимальное значение метрики BLEU составило 13.

3.3. Использование корпуса «Еврорадио» и предобработка данных

Следующим этапом к кодексам РБ был добавлен параллельный корпус «Еврорадио» и появилась первая предобработка данных. Объём данных для обучения – не единственный показатель успешных результатов. Немало зависит и от их качества. В опыте, проведённом в части два главы три, никакие преобразования исходных данных не производились. Для нейронной сети слова «Паспорт» и «паспорт» не имеют ничего общего, поэтому, если сеть

знает перевод для слова «паспорт», то же слово, но написанное с заглавной буквы, она оставит без перевода. Поэтому в первую очередь приведём все буквы в тексте к одному регистру. Числа остаются неизменными в любом языке – уберём их из обучения. Так же уберём знаки пунктуации и специальные символы, а также пустые строки – пунктуацию вернём в будущем, а пустые строки не имеет смысла переводить. В итоговом корпусе не будет букв и слов из других языков.

Для выполнения всех указанных выше преобразований был написан скрипт `clear\_data.py`, который получает на вход файл, требующий коррекции, и создаёт новый документ с чистым форматом, пригодным к использованию нейронной сетью. Результаты работы скрипта приведены в таблице 3.2.

<i>исходный</i>	Воздушный кодекс Республики Беларусь устанавливает правовые и организационные основы использования воздушного пространства Республики Беларусь и осуществления деятельности в области авиации в целях обеспечения потребностей граждан и экономики, обороны и безопасности государства.
<i>преобразованный</i>	воздушный кодекс республики беларусь устанавливает правовые и организационные основы использования воздушного пространства республики беларусь и осуществления деятельности в области авиации в целях обеспечения потребностей граждан и экономики обороны и безопасности государства
<i>исходный</i>	ГЛАВА 1
<i>преобразованный</i>	Глава
<i>исходный</i>	Статья 1. Основные термины и их определения, применяемые в настоящем Кодексе
<i>преобразованный</i>	статья основные термины и их определения применяемые в настоящем кодексе

Таблица 3.2 – Результаты работы `clear\_data.py` с воздушным кодексом РБ.

Из таблицы 3.2 видно, что в результате приведения к одному регистру пострадало название, которое всегда пишется с использованием заглавных букв – «Республика Беларусь». Это даёт основание для последующего улучшения алгоритма – в конечном результате имена собственные не должны быть написаны некорректно.

Отметим также разделение корпуса на тренировочные и проверочные данные: 91590 строк для обучения и около 45910 для этапа валидации. Таким

образом, данные поделены в пропорциях 70/30. Тестирование будем проводить на оставшихся параллельных кодексах.

В результате обучения в 100000 шагов на видеокарте GeForce GTX 850 М по прошествии суток был получен связный результат. В данных не было чисел и пунктуации, но переведённый текст оставался связным и читабельным.

3.4. Нормализация текста

Несмотря на связность перевода, отсутствие чисел, как и спецсимволов и знаков пунктуации, делает перевод малопригодным в повседневной жизни. Для исправления этого недостатка, а также для сохранения символов, не подлежащих переводу, был подготовлен код с использованием библиотеки `stanfordnlp` (в последствии - `stanza`). Библиотека даёт возможность приводить слова к корректному регистру. Однако её подход не совершенен: в некоторых случаях при склонении «Республика Беларусь» превращается в «республика Беларусь», так как слово «республика» чаще употребляется в качестве нарицательного. Видно, что сила контекста в данном случае задействована не полностью. Так же стоит отметить отсутствие пакета для белорусского языка (он был добавлен позже и внедрён в более поздние версии переводчика). Из-за чего белорусская буква «і» обрабатывалась некорректно и отделялась от слов. В дальнейшем проблема была исправлена. Код для преобразования исходного текста в отформатированный имеет следующий вид:

```
import re
import stanfordnlp

stf_nlp = stanfordnlp.Pipeline(processors='tokenize,mwt,pos', lang='ru')

def preprocess_text(text):
    new_text = []

    for i in range(len(text)):
        t_ = text[i].replace("\xa0", " ")
        t_ = ' '.join(re.findall(r"[\w']+|[/.\s*!\"'?:;%<>:№-]", t_))
        if not t_:
            continue
        doc = stf_nlp(t_)

        words = []
        for sent in doc.sentences:
            for w in sent.words:
                words.append(w.text.capitalize()
                              if w.upos in ["PROPN", "NNS", "NNPS"]
                              else w.text.lower())

        if words:
            new_text.append(' '.join(words) + '\n')
    return new_text
```

В результате использования этого кода:

- происходит разделение знаков пунктуации и других знаков от слов;
- знак «нераздельного пробела» заменяется на обычный пробел;
- слова приводятся к соответствующему им регистру;
- убираются пустые строки.

Пример обработки текста приведён в таблице 3.3.

<i>исходный</i>	Воздушный кодекс Республики Беларусь устанавливает правовые и организационные основы использования воздушного пространства Республики Беларусь и осуществления деятельности в области авиации в целях обеспечения потребностей граждан и экономики, обороны и безопасности государства.
<i>преобразованный</i>	воздушный кодекс Республики Беларусь устанавливает правовые и организационные основы использования воздушного пространства Республики Беларусь и осуществления деятельности в области авиации в целях обеспечения потребностей граждан и экономики , обороны и безопасности государства .
<i>исходный</i>	ГЛАВА 1
<i>преобразованный</i>	глава 1
<i>исходный</i>	Статья 1. Основные термины и их определения, применяемые в настоящем Кодексе
<i>преобразованный</i>	статья 1 . основные термины и их определения , применяемые в настоящем Кодексе

Таблица 3.3 – Результаты предобработки воздушного кодекса РБ.

Такой обработки достаточно для обучения и проверки перевода: знаки пунктуации отделены от слов, чтобы одно и то же слово не воспринималось сетью по-разному.

Отметим также, что для данного эксперимента данные для обучения и валидации были перераспределены: теперь нейросеть обучается на 132089 строках текста, в котором наибольшую часть составляют данные «Еврорадио», но также присутствует часть кодекса о земле и полный лесной кодекс, в валидационную часть попало порядка 5400 строк, в который вошёл остаток корпуса «Еврорадио» и оставшаяся часть кодекса о земле.

Обученной нейронной сети было предложено перевести с русского языка на белорусский несколько вариантов текста: «Кодекс РБ о браке и семье», «Воздушный кодекс РБ» и «Война и мир» Льва Николаевича Толстого. Выдержки из полученных переводов приведены в таблицах 3.4-6 соответственно.

Язык	Текст
русский	документы , выданные компетентными органами иностранных государств в удостоверение актов гражданского состояния , совершенных за пределами Республики Беларусь с соблюдением законодательства соответствующих государств в отношении граждан Республики Беларусь , иностранных граждан и лиц без гражданства , признаются действительными в Республике Беларусь при наличии консульской легализации , если иное не предусмотрено законодательством Республики Беларусь , в том числе международными договорами Республики Беларусь .
белорусский (NN)	дакументы , выдадзеныя кампетэнтнымі organамі замежных дзяржаў у пасведчанне актаў <i>грамадзянскай</i> стану , учыненых за <u>межам і Рэспублік і</u> Беларусь з захаваннем заканадаўства адпаведных дзяржаў у дачыненні да грамадзян <u>Рэспублік і</u> Беларусь , замежных грамадзян і асобаў без грамадзянства , прызнаюцца <i>Бутквяічус</i> у Рэспубліцы Беларусь пры наяўнасці <u>консульскай легалізацы і</u> , калі іншае не прадугледжана заканадаўствам Рэспублікі Беларусь , у тым ліку міжнароднымі <u>договорами Ў рэспублікі</u> Беларусь .
русский	привести решения Правительства Республики Беларусь в соответствие с настоящим Кодексом ;
белорусский (NN)	прывесці рашэнні <i>Урада</i> Рэспублікі Беларусь у адпаведнасць з гэтым Кодэксам ;

Таблица 3.4 – Результаты перевода «Кодекс РБ о браке и семье» с помощью обученной сети.

Как видно из таблицы 3.4, сеть распознаёт большее количество слов, текст стал более связным и обрёл смысл. Проблемы возникают с формами слов «договор» - «договоры» и их склонения, а также с окончанием «і», которое было некорректно обработано при предварительном изменении текста. Так как преобразование осуществлялось в соответствии с русскими правилами, то белорусские буквы и апострофы, не встречающиеся в русских словах, вызвали некорректные результаты. Эти недочёты несложно исправить, поэтому при следующем эксперименте они будут вынесены.

В результатах, приведённых в таблице 3.5, прослеживаются трудности обученной сети при взаимодействии с узкоспециализированной лексикой (например, слово «неподвижных»). Так же в сравнении с человеческим переводом можно заметить вариативность перевода некоторых отдельных слов или некорректно подобранное окончание.

Язык	Текст
<i>русский</i>	маркировка неподвижных объектов и сооружений дневными и ночными маркировочными знаками , устройствами и отметками производится владельцами этих объектов и сооружений за счет собственных средств .
<i>белорусский (NN)</i>	маркіроўка <u>неподвижных</u> <u>аб'ектаў</u> і збудаванняў <u>дневными</u> і <u>ночным</u> і <u>маркировочными</u> <u>знакамі</u> , <u>устройствами</u> і <u>отметками</u> <u>робіцца</u> ўладальнікамі гэтых аб'ектаў і збудаванняў за кошт уласных сродкаў .
<i>белорусский (профессиональный)</i>	маркіроўка нерухомах аб'ектаў і збудаванняў дзённымі і начнымі <u>маркіровачнымі</u> <u>знакамі</u> <u>ладамі</u> і <u>адзнакамі</u> <u>праводзіцца</u> ўладальнікамі гэтых аб'ектаў і збудаванняў за кошт уласных сродкаў
<i>русский</i>	принимать на территории Республики Беларусь на борт воздушных судов пассажиров , багаж , грузы и почтовые отправления для воздушных перевозок в пределах территории Республики Беларусь по разрешению специально уполномоченного органа в области гражданской авиации .
<i>белорусский (NN)</i>	прымаць на <u>тэрыторыі</u> і <u>рэспублік</u> і Беларусь на борт паветраных <u>судоў</u> <u>пасажыра</u> <u>ў</u> , багаж , грузы і паштовыя адпраўленні для паветраных перавозак у <u>межах</u> тэрыторыі Рэспублікі Беларусь па дазволе спецыяльна ўпаўнаважанага органа <u>ў</u> <u>галіне</u> грамадзянскай <u>авіяцыі</u> і .
<i>белорусский (профессиональный)</i>	прымаць на тэрыторыі рэспублікі беларусь на борт паветраных <u>судоў</u> <u>пасажыраў</u> багаж грузы і паштовыя адпраўленні для паветраных перавозак у <u>граніцах</u> тэрыторыі рэспублікі беларусь па дазволе спецыяльна ўпаўнаважанага органа ў вобласці грамадзянскай <u>авіяцыі</u>

Таблица 3.5 – Результаты перевода «Воздушный кодекс РБ» с помощью обученной сети с сохранением пунктуации.

Результаты, приведённые в таблице 3.6, демонстрируют возможности NN при столкновении с произведением Л.Н. Толстого. «Война и мир» - художественное произведение, изобилующее словами и выражениями, не используемыми в современном общении, поэтому подготовленный нами

корпус не справляется со словарём Льва Николаевича Толстого и перевод успешным считать нельзя. Однако он демонстрирует ещё один недостаток текущего обучения: «шел оживленный разговор» переводится как «ішоў *оживленный разговор*», то есть перевод предложен и он корректен, но ошибка в падеже.

Язык	Текст
<i>русский</i>	вечером того же дня на квартире Денисова шел оживленный разговор офицеров эскадрона .
<i>белорусский (NN)</i>	увечары таго ж дня на кватэры Дзянісава ішоў <i>оживленный разговор</i> афіцэраў эскадрона .
<i>русский</i>	вошел полковой адъютант и подтвердил известие , привезенное Жерковым . назавтра велено было выступать .
<i>белорусский (NN)</i>	увайшоў <i>полковой Паўліка і пацвердзіў Судана , привезенное Жерковым .</i> назаўтра велено было выступаць .

Таблица 3.6 – Результаты перевода «Война и мир» с помощью сети.

Как упоминалось в части первой этой главы, оценку качества будем производить на «Воздушном кодексе РБ», в качестве метрики будем использовать BLEU, представленный в пункте 1 главы 1.

Так как в BLEU используются n -граммы слов, при $n \in \{1, 2, 3, 4\}$, с разными весами этих n -грамм, то и BLEU может варьироваться.

1-gramm	2-gramm	3-gramm	4-gramm	BLEU: 1	BLEU: 2
0.25	0.25	0.25	0.25	38.6	46.6
0	0	0	1	25.3	34.8
0	0	1	0	34.7	45.3
1	0	0	0	69.7	76.6
0	0	0.5	0.5	28.9	38.4
0.75	0.25	0	0	62.3	71.4
0	0.33	0.33	0.34	33.0	42.2
auto	auto	auto	auto	41.1	49.9

Таблица 3.7 – Значения BLEU метрики в зависимости от весов n -грамм, используемых для получения результата, а также номера эксперимента, различающихся процессом подготовки данных для обучения и перевода.

В данном случае, для подсчёта BLEU использовалась встроенная функция библиотеки nltk «sentence_bleu». Результаты метрики для разных весов приведены в таблице 3.7. Колонка «BLEU 1» демонстрирует значение метрики для эксперимента с данными, полностью приведёнными к нижнему регистру, с отсутствием цифр и пунктуации, колонка «BLEU 2» отвечает за результаты метрики в случае использования библиотеки StanfordNLP и дополнительных преобразований, описанных выше.

Как видно из таблицы 3.7, более крупные n-граммы дают более низкое значение метрики, при учёте только униграмм результат наиболее высокий. Так же стоит отметить, что значения метрики для второго эксперимента ощутимо выше, чем для первого. Одним из объяснений такого результата может служить присутствие спецсимволов, цифр и других знаков, которые увеличивают общую длину предложений и влияют на количество совпадающих n-грамм в BLEU 2.

3.5. Работа с редкими словами: добавление SentencePiece

В последнем эксперименте наиболее остро прослеживается невозможность перевода незнакомых NN слов, поэтому следующим шагом стало изменение предобработки данных в сторону дробления слов. Как уже отмечалось выше, удобный интерфейс предоставляет библиотека SentencePiece. Пример её работы можно увидеть в таблице 3.8.

Язык	Текст
<i>русский</i>	вечером того же дня на квартире Денисова шел оживленный разговор офицеров эскадрона как-нибудь.
<i>русский - результат</i>	__вечером __того __же __дня __на __квартире __Денисов а __шел __оживленн ый __разговор __офицеров __э ска др она __как - нибудь .
<i>белорусский</i>	голас яго пасля сну быў нямоцны, высільвацца яму не хацелася, і ён як бы памагаў сабе лянівым, але звонкім ляскам пугі.
<i>белорусский - результат</i>	__голас __яго __пасля __сну __быў __ня моцны , __вы сі ль вацца __яму __не __хацелася , __і __ён __як __бы __па м ага ў __сабе __ля ні вым , __але __з во нкі м __ля ска м __пу гі .

Таблица 3.8 – Пример обработки текстов на русском и белорусском языках с помощью возможностей SentencePiece. Тип модели – unigram, размер словаря – 32000.

Отметим, что знаки препинания и частицы слов здесь разделены, однако библиотека позволяет производить декодирование в обратном порядке,

восстанавливая исходный текст. Такой подход даёт удобство в работе с пунктуацией и другими специальными символами.

Для экспериментов были рассмотрены три вариации размера словаря: 32000, 16000 и 4000. Несмотря на то, что в исходном описании библиотеки для малобуквенных языков качество перевода улучшалось при снижении размера словаря, текущие эксперименты не позволяют однозначно подтвердить такое заявление. На практике оказалось, что, действительно, качество перевода лучше для словаря меньшей размерности на предложениях небольшой длины, но при увеличении количества слов в предложении происходит обрезка и/или зацикливание, что существенно хуже. Поэтому для более корректного перевода с помощью словаря на 4000 частиц необходимо проводить дополнительное разделение длинных предложений на отдельные части и рассматривать их как независимые предложения.

Код для создания модели SentencePiece представлен ниже.

```
import os
import sentencepiece as spm

def train_model(input_file, model_prefix='demo', vocab_size=2000,
model_type='unigram',):
    filepath =
os.path.join(r"C:\Users\MY\Documents\data\n\projects\translator_preprocessing\tran
slator_preprocessing\notebooks\data\splited", input_file)
    spm.SentencePieceTrainer.train(f'--input="{filepath}" --
model_prefix={model_prefix} --vocab_size={vocab_size} --model_type={model_type}')

    sp = spm.SentencePieceProcessor()
    sp.load(f'{model_prefix}.model')
    return sp
```

Разработчики данной библиотеки так же уверяют, что для работы вполне подойдут сырые данные, которые не проходили стадию обработки. Однако эти слова не стоит распространять на следующую за библиотекой нейронную сеть: для её обучения данные всё так же требуются в обработке, в связи с чем для достижения лучшего результата необходимо предварительно обработать данные. Для обработки случая с текстом «Республика Беларусь», в котором слово «Республика» в некоторых случаях преобразовывалось к нижнему регистру, были введены дополнительные ограничения при предобработке данных.

3.6. Результаты экспериментов

Для удобства сравнения, из описанных в таблице 3.7 вариаций метрики BLEU, оставим 3: 1-gramm, 4-gramm, 0.25 для каждой из четырёх. Результаты метрик приведены в таблице 3.9.

Из таблицы видно, что наилучший результат достигается при наибольшем количестве шагов обучения. Однако при фиксированном количестве шагов и вариации словаря или вида входных данных, победа не так очевидна.

№	Эксперимент	0.25 для каждой	только 4-gramm	только 1-gramm
3	Словарь – 32000. Сырые данные без предобработки. 50000 шагов.	18.5	17.2	27.4
4	Словарь – 16000. Сырые данные без предобработки. 50000 шагов.	10.6	8.6	22.6
5	Словарь – 32000. Данные предобработаны. 50000 шагов.	<u>49.5</u>	38.8	<u>76.3</u>
6	Словарь – 16000. Данные предобработаны. 50000 шагов.	48.8	<u>38.9</u>	75.1
7	Словарь – 32000. Данные предобработаны. «Одна строка – одно предложение» 50000 шагов.	48.6	38.1	75.6
8	Словарь – 32000. Данные предобработаны. 150000 шагов.	52.4	41.6	<u>78.6</u>
9	Словарь – 4000. Данные предобработаны. 150000 шагов.	<u>53.4</u>	<u>44.5</u>	76.6
10	Словарь – 4000. Данные предобработаны. «Одна строка – одно предложение» 150000 шагов.	53.1	44.0	77.3

Таблица 3.9 – Метрика BLEU для экспериментов с библиотекой SentencePiece.

Для лучшего понимания результатов, в таблицах 3.10-3.12 приведены переводы одинаковых фрагментов сетями, полученными в ходе разных экспериментов.

№	Текст
оригинал	Статья 169. Опекунны, попечители детей-сирот, детей, оставшихся без попечения родителей, и совершеннолетних подопечных Опекунами, попечителями детей-сирот, детей, оставшихся без попечения родителей, устроенных на воспитание в приемные семьи, детские дома семейного типа, являются приемные родители, родители-воспитатели детского дома семейного типа..
3	Артыкул 169. <i>Апетуны, папечылі</i> дзяцей-сірот, дзяцей, якія засталіся без апекі бацькоў, і <i>непаўналетніх</i> падапечных <i>Аспеткамі, папечымі</i> дзяцей-сірот, дзяцей, якія засталіся без апекі бацькоў, <i>убуданых</i> на выхаванне ў прыёмныя сем’і, дзіцячыя дамы сямейнага тыпу, з’яўляюцца прыёмныя бацькі, бацькі- <i>аўсхальнікі</i> дзіцячага дома сямейнага тыпу.
5	артыкул 169. апекуны, <i>попелі</i> дзяцей-сірот, дзяцей, якія засталіся без апекі бацькоў, і <i>непаўналетніх</i> падапечных апекунамі, <i>папячычным</i> дзяцей-сірот, дзяцей, якія засталіся без апекі бацькоў, <i>убудавааных</i> на выхаванне ў прыёмныя сем’і, дзіцячыя дамы сямейнага тыпу, з’яўляюцца прыёмныя бацькі, бацькі- <i>выхасталі</i> дзіцячага дома сямейнага тыпу.
8	артыкул 169. <i>апякуны, паданні</i> дзяцей-сірот, дзяцей, якія засталіся без апекі бацькоў, і <i>паўналетні</i> падапечных <i>апякушамі, папечамі</i> дзяцей-сірот, дзяцей, якія засталіся без апекі бацькоў, <i>зладжаных</i> на выхаванне ў прыёмныя сем’і, дзіцячыя дамы сямейнага тыпу, з’яўляюцца прыёмныя бацькі, <i>то</i> бацькі- <i>ввотнасці</i> дзіцячага дома сямейнага тыпу.
9	артыкул 169. апекуны, <i>апеллі</i> дзяцей-сірот, дзяцей, якія засталіся без апекі бацькоў, і <i>зусімгадовых</i> падапечных надапечных апекунамі, <i>спячільнікамі</i> дзяцей-сірот, дзяцей, якія засталіся без апекі бацькоў, <i>уладкаваных</i> на выхаванне ў прыёмныя сем’і, дзіцячыя дамы сямейнага тыпу, з’яўляюцца прыёмныя бацькі, бацькі- <i>выхаваўцы</i> дзіцячага дома сямейнага тыпу.
10	артыкул 169. апекуны, <i>папялі</i> дзяцей-сірот, дзяцей, якія засталіся без апекі бацькоў, і <i>зусімгадовых</i> падапечных надапечных надапечных апекунамі, <i>спячільнікамі</i> дзяцей-сірот, дзяцей, якія засталіся без апекі бацькоў, <i>уладкаваных</i> на выхаванне ў прыёмныя сем’і, дзіцячыя дамы сямейнага тыпу, з’яўляюцца прыёмныя бацькі, бацькі- <i>выхаваўцы</i> дзіцячага дома сямейнага тыпу.

Таблица 3.10 – Примеры перевода текста с русского языка на белорусский NN из разных экспериментов. Источник: «Кодекс РБ о браке и семье».

№	Текст
оригинал	Документы, выданные компетентными органами иностранных государств в удостоверение актов гражданского состояния, совершенных за пределами Республики Беларусь с соблюдением законодательства соответствующих государств в отношении граждан Республики Беларусь, иностранных граждан и лиц без гражданства, признаются действительными в Республике Беларусь при наличии консульской легализации, если иное не предусмотрено законодательством Республики Беларусь, в том числе международными договорами Республики Беларусь.
2	дакументы , выдадзеныя кампетэнтнымі organамі замежных дзяржаў у пасведчанне актаў <i>грамадзянскай</i> стану , учыненых за межамі Рэспублікі і Беларусь з захаваннем заканадаўства адпаведных дзяржаў у дачыненні да грамадзян Рэспублікі і Беларусь , замежных грамадзян і асобаў без грамадзянства , прызнаюцца <i>Буткявічус</i> у Рэспубліцы Беларусь пры наяўнасці <i>консульскай</i> легалізацыі і , калі іншае не прадугледжана заканадаўствам Рэспублікі Беларусь , у тым ліку міжнароднымі <i>дагаворами</i> Ў рэспублікі Беларусь .
3	Дакументы, выдадзеныя кампетэнтнымі organамі замежных дзяржаў у пасведчанне актаў <i>грамадзянскай</i> стану, учыненых за межамі Рэспублікі Беларусь з захаваннем заканадаўствам адпаведных дзяржаў у дачыненні да грамадзян Рэспублікі Беларусь, замежных грамадзян і асобаў без грамадзянства, прызнаюцца сапраўднымі ў Рэспубліцы Беларусь пры наяўнасці консульскай легалізацыі, калі іншае не прадугледжана заканадаўствам Рэспублікі Беларусь, у тым ліку міжнароднымі <i>дагаворами</i> Рэспублікі Беларусь.
5	дакументы, выдадзеныя кампетэнтнымі organамі замежных <i>дзяржаваў</i> у пасведчанне актаў <i>грамадзянскай</i> стану, здзейсненых за межамі Рэспублікі Беларусь з захаваннем заканадаўства адпаведных дзяржаў у дачыненні да грамадзян Рэспублікі Беларусь, <i>у тым ліку міжнароднымі дамовы Беларусь, у тым ліку міжнароднымі дамовы Беларусь, у тым ліку міжнароднымі дамовы Беларусь, у тым ліку міжнароднымі дамовы Беларусь.</i>
8	дакументы, выдадзеныя кампетэнтнымі organамі замежных дзяржаў у пасведчанне актаў <i>грамадзянскай</i> стану, учыненых за межамі Рэспублікі Беларусь з выкананнем заканадаўства адпаведных дзяржаў у дачыненні да грамадзян Рэспублікі Беларусь, замежных грамадзян і асобаў без грамадзянства, прызнаюцца сапраўднымі ў Рэспубліцы Беларусь пры наяўнасці консульскай легалізацыі, калі іншае не прадугледжана заканадаўствам Рэспублікі Беларусь, у тым ліку міжнароднымі <i>дамовымі</i> Рэспублікі Беларусь.

Таблица 3.11 – Примеры перевода текста с русского языка на белорусский NN из разных экспериментов. Источник: «Кодекс РБ о браке и семье»

Из таблиц видно, что все NN стараются перевести неизвестные слова. В некоторых случаях удаётся очень близко воспроизвести слово, в других же слова абсолютно далеки от реальности. Отметим успехи в переводе сложных страдательных деепричастий: «оставшихся» переходит в «якія засталіся».

Посмотрим более внимательно на сравнение перевода, выполненного NN, с человеческим переводом.

№	Текст
оригинал	маркіроўка нерухомах аб'ектаў і збудаванняў дзённымі і начнымі маркіровачнымі знакамі, ладамі і адзнакамі праводзіцца ўладальнікамі гэтых аб'ектаў і збудаванняў за кошт уласных сродкаў.
2	маркіроўка <i>неподвижных</i> аб'ектаў і збудаванняў <i>дневными і ночнымі маркіровачнымі</i> знакамі , <i>устройствами і отметками</i> робіцца ўладальнікамі гэтых аб'ектаў і збудаванняў за кошт уласных сродкаў .
5	<i>маркіроўку непаруных</i> аб'ектаў і збудаванняў <i>дзённымі і навінамі маркічкамі</i> знакамі, прыладамі і адзнакамі вырабляецца ўладальнікам гэтых аб'ектаў і збудаванняў за кошт уласных сродкаў.
7	маркіроўка <i>непавітаных</i> аб'ектаў і <i>пабудовамі здольнасць і начным марківачамі</i> знакамі, прыладамі і адзнакамі вырабляецца ўладальнікамі гэтых аб'ектаў і збудаванняў за кошт уласных сродкаў.
8	маркіроўка <i>нерухвальных</i> аб'ектаў і збудаванняў <i>дзённымі інснымі марківачнымі</i> знакамі, прыладамі і адзнакамі вырабляецца ўладальнікамі гэтых аб'ектаў і збудаванняў за кошт уласных сродкаў.
оригинал	прымаць на тэрыторыі Рэспублікі Беларусь на борт паветраных судоў пасажыраў, багаж, грузы і паштовыя адпраўленні для паветраных перавозак у граніцах тэрыторыі Рэспублікі Беларусь па дазvole спецыяльна ўпаўнаважанага органа ў вобласці грамадзянскай авіяцыі.
2	прымаць на тэрыторыі і рэспублік і Беларусь на борт паветраных судоў пасажыраў <i>Ў</i> , багаж , грузы і паштовыя адпраўленні для паветраных перавозак у межах тэрыторыі Рэспублікі Беларусь па дазvole спецыяльна ўпаўнаважанага органа <i>Ў</i> галіне грамадзянскай авіяцыі і .
5	прымаць на тэрыторыі Рэспублікі Беларусь на борт паветраных судоў пасажыраў, багаж, грузы і паштовыя адпраўленні для паветраных перавозак у граніцах тэрыторыі Рэспублікі Беларусь па развязе спецыяльна ўпаўнаважанага органа ў галіне грамадзянскай авіяцыі.

Таблица 3.12 – Примеры перевода текста с русского языка на белорусский NN из разных экспериментов. Сравнение с переводом человека.

7	прымаць на тэрыторыі Рэспублікі Беларусь на борт паветраных <i>суднах</i> пасажыраў, багаж, грузы і паштовыя адпраўленні для паветраных перавозак у межах тэрыторыі Рэспублікі Беларусь па дазvole спецыяльна ўпаўнаважанага органа ў галіне грамадзянскай авіяцыі.
8	прымаць на тэрыторыі Рэспублікі Беларусь на борт паветраных <i>судоў</i> пасажыраў, багаж, грузы і паштовыя адпраўленні для паветраных перавозак у межах тэрыторыі Рэспублікі Беларусь па <i>развязе</i> спецыяльна ўпаўнаважанага органа ў галіне грамадзянскай авіяцыі.
10	прымаць на тэрыторыі Рэспублікі Беларусь на борт паветраных <i>судоў</i> пасажыраў, багаж, грузы і паштовыя адпраўленні для паветраных перавозак у межах тэрыторыі Рэспублікі Беларусь <u>па дазvole грамадзянскай авіяцыі</u> .

Таблица 3.12 – Примеры перевода текста с русского языка на белорусский NN из разных экспериментов. Сравнение с переводом человека. Продолжение.

В приведённой таблице бросается в глаза поведение последней обученной сети: небольшой размер словаря подстрекает NN к расхождению со структурой предложения, то добавляя повторения, то вырезая целые части. Несмотря на то, что перевод отдельных неизвестных слов ощутимо улучшился, связность некоторых предложений пострадала. Как и в таблице выше, показано, как NN предлагают свои переводы неизвестных слов. Для улучшения качества предлагаемых вариантов необходимо улучшить качество частиц, на которые делятся слова. Это возможно совершить с помощью увеличения размера корпуса, на котором происходит обучение модели. Как уже отмечалось ранее, тексты для разных языков не обязаны быть параллельны, что упрощает поиск данных для обучения.

Стоит так же отметить временные затраты: помимо обучения самой NN, существенное время уходит на предварительную обработку данных, в частности, на определение части речи с помощью StanfordNLP/stanza. На преобразование файла с тренировочными данными было потрачено 2.5 часа. Конечно, время обработки зависит от мощности видеокарты, однако при малых её ресурсах, время обработки оказывается очень существенным.

3.7. Пользовательская оболочка

Для демонстрации результатов работы был написан сервис-переводчик для интеграции в интернет-портал «corpus.by». Вид сервиса представлен на рисунке 3.1. Сервис состоит из двух текстовых полей, одно для ввода текста на русском языке, другое – для вывода результатов перевода. Осуществить

перевод можно по нажатию кнопки «Перакладчы». Со стороны сервиса будет произведён весь цикл подготовки текста к переводу: приведение к корректному регистру, разделение абзацев на отдельные предложения, разделение слов на более мелкие частицы. Далее последует непосредственно сам перевод с помощью предобученной NN, завершающим этапом идёт обратное объединение частиц в слова, обособленных предложении – в абзацы. Пункт приведения к исходному регистру не производится.

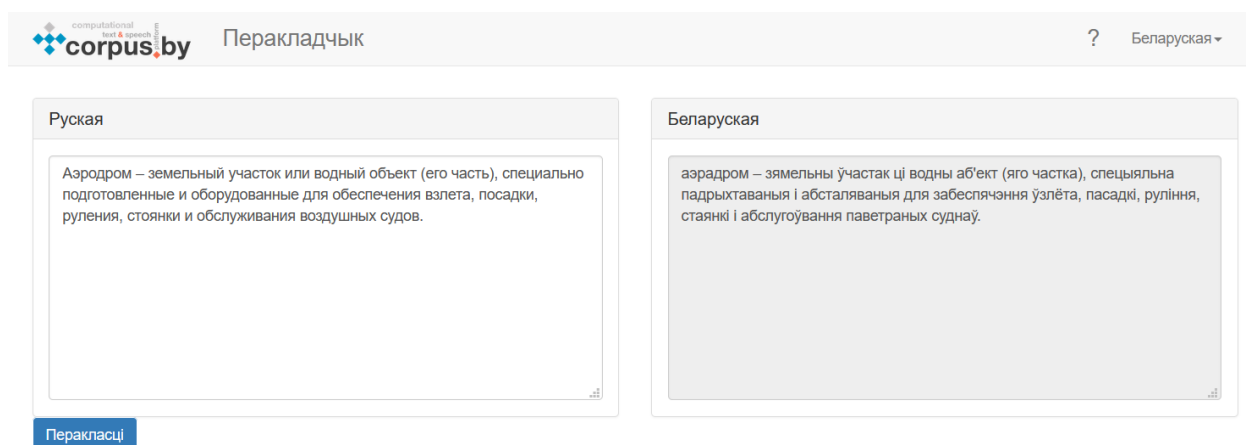


Рисунок 3.1. Сервис «Переводчик» на corpus.by.

Благодаря широкому набору сервисов, которые представлены на сайте, для пользователя переводчика, описанного в работе, подготовлен удобный интерфейс, а также локализация на трёх языках.

3.8. Выводы

Во третьей главе получены следующие основные результаты:

- описан программный эксперимент, в ходе которого были описаны все шаги построения переводчика юридических текстов с русского на белорусский язык;
- описан этап поиска данных для обучения, а также этап подготовки данных с помощью различных библиотек;
- продемонстрированы промежуточные оценки метрики BLEU;
- показаны результаты работы обученных сетей на текстах юридической и не юридической тематики, выполнен подсчёт метрики качества перевода;
- на основании проведённых экспериментов выявлены слабые и сильные стороны текущего решения;

- описан пользовательский сервис для взаимодействия с итоговой моделью переводчика;
- интегрирован сервис перевода в экосистему сервисов «corpus.by».

ЗАКЛЮЧЕНИЕ

В ходе проведённых исследований области были выявлены существующие решения, проанализированы их достоинства и недостатки и определены инструменты, которые используются для переводов между естественными языками. Также были выявлены и описаны существующие подходы к решению задачи, изучены их алгоритмы.

Важной частью работы стали поиск и подготовка параллельного корпуса для перевода, были изучены и опробованы различные подходы, проведён сравнительный анализ и выбран наилучший для достижения цели работы.

Стоит так же отметить, что для демонстрации работы переводчика и получения обратной связи был подготовлен сервис для сайта лаборатории распознавания и синтеза речи национальной академии наук Республики Беларусь. Текущая версия сервиса предлагает перевод со всем циклом обработки исходного текста. Его присутствие на данном портале позволяет расширить пользовательскую аудиторию переводчика.

На основании проделанной работы можно сказать, что для достижения поставленной цели были важны исходные данные, а именно – параллельный корпус текста. Значимой частью работы так же является предобработка данных, для корректного перевода слов, встречающихся в тексте в разных регистрах и падежах, а также производные от них. Обученная сеть также показала способность конструировать возможный перевод из уже известных ей параметров (переводов корней, суффиксов и тому подобных частей слова), однако он, хоть и близок, всё же не точный, поэтому требует большего количества исходных данных для повышения качества предлагаемых вариантов.

Также отметим, что качество перевода предложений средней длины с помощью обученной модели значительно лучше, чем для предложений очень коротких или очень длинных. Данная особенность так же может быть решена с помощью расширения набора текстов для обучения.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. OpenNMT-py: Open-Source Neural Machine Translation // [Электронный ресурс]. — . Режим доступа : <https://github.com/OpenNMT/OpenNMT-py>.
2. BLEU: a Method for Automatic Evaluation of Machine Translation // aclweb [Электронный ресурс]. — 2002. Режим доступа : <https://www.aclweb.org/anthology/P02-1040>. — Дата доступа : 26.05.2019.
3. ОСОБЕННОСТИ ЮРИДИЧЕСКОГО ПЕРЕВОДА // [Электронный ресурс]. — 2018. Режим доступа : <http://lingvotech.com/uridicheskiper>. — Дата доступа : 11.11.2018.
4. Особенности юридического перевода // [Электронный ресурс]. — 2018. Режим доступа : http://www.norma-tm.ru/osobjennosti_juridichjeskogo_pjerjevoda.html. — Дата доступа : 11.11.2018.
5. Перевод с белорусского языка // [Электронный ресурс]. — 2018. Режим доступа : http://www.primavista.ru/rus/example/e_belorus. — Дата доступа : 11.11.2018.
6. ОСОБЕННОСТИ ПЕРЕВОДА БЕЛОРУССКОГО ЯЗЫКА // [Электронный ресурс]. — 2018. Режим доступа : <http://word-house.ru/yazyki-perevoda/beloruskiy-yazyk/osobennosti-perevoda-beloruskogo-yazyka/>. — Дата доступа : 11.11.2018.
7. Google Translator // [Электронный ресурс]. — 2018. Режим доступа : <https://translate.google.com/>.
8. Яндекс.Переводчик // [Электронный ресурс]. — 2018. Режим доступа : <https://translate.yandex.by/>.
9. Анлайн-перакладнік Белазар // [Электронный ресурс]. — 2018. Режим доступа : <http://belazar.info/>.
10. ЮРИДИЧЕСКИЙ СЛОВАРЬ // [Электронный ресурс]. — 2018. Режим доступа : <http://multilang.etalonline.by/>. — Дата доступа : 11.11.2018.
11. NooJ: A Linguistic Development Environment // [Электронный ресурс]. — 2018. Режим доступа : <http://www.nooj-association.org/>.
12. Translate - a PyTorch Language Library // [Электронный ресурс]. — 2018. Режим доступа : <https://github.com/pytorch/translate>.
13. Fairseq(-py) // [Электронный ресурс]. — 2018. Режим доступа : <https://github.com/pytorch/fairseq>.
14. Towards Zero Unknown Word in Neural Machine Translation // www.nlpr.ia.ac.cn [Электронный ресурс]. — 2016. Режим доступа : http://www.nlpr.ia.ac.cn/cip/ZongPublications/2016/IJCAI_2016_LXQ.pdf. — Дата доступа : 10.05.2020.
15. Sequence to Sequence Learning with Neural Networks // [Электронный ресурс]. — . Режим доступа : <https://papers.nips.cc/paper/5346-sequence-to-sequence-learning-with-neural-networks.pdf>.
16. A Convolutional Encoder Model for Neural Machine Translation // [Электронный ресурс]. — . Режим доступа : <http://aclweb.org/anthology/P/P17/P17-1012.pdf>.
17. SentencePiece // [github.com](https://github.com/google/sentencepiece) [Электронный ресурс]. — 2020. Режим доступа : <https://github.com/google/sentencepiece>. — Дата доступа : 16.04.2020.

18. SentencePiece // [github.com](https://github.com/google/sentencepiece/blob/master/doc/experiments.md#results-bleu-scores-1) [Электронный ресурс]. – 2020. Режим доступа : <https://github.com/google/sentencepiece/blob/master/doc/experiments.md#results-bleu-scores-1>. – Дата доступа : 16.04.2020.
19. Recurrent neural network // [Электронный ресурс]. — . Режим доступа : https://en.wikipedia.org/wiki/Recurrent_neural_network.
20. OpenNMT: Neural Machine Translation Toolkit // [arxiv.org](https://arxiv.org/pdf/1805.11462.pdf) [Электронный ресурс]. – 2018. Режим доступа : <https://arxiv.org/pdf/1805.11462.pdf>. – Дата доступа : 07.01.2019.
21. Аўтаматычны перакладчык з рускай на беларускую // [euroradio.fm](https://euroradio.fm/сyаpеr-z-nulyа-mozhna-lyogkа-zrаbіс-autаmаtусhny-pеrаklаdсhуk-z-ruskау-nа-bеlаruskuyu) [Электронный ресурс]. – 2018. Режим доступа : <https://euroradio.fm/сyаpеr-z-nulyа-mozhna-lyogkа-zrаbіс-autаmаtусhny-pеrаklаdсhуk-z-ruskау-nа-bеlаruskuyu>. – Дата доступа : 20.05.2019.