ПРАВИТЕЛЬСТВО РОССИЙСКОЙ ФЕДЕРАЦИИ
ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ
ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ
«ВЫСШАЯ ШКОЛА ЭКОНОМИКИ»

Факультет компьютерных наук
Образовательная программа «Программная инженерия»

УДК 004.8

СОГЛАСОВАНО
Руководитель проекта,
доцент департамента анализа данных и
искусственного интеллекта,
канд. техн. наук

_____ Д.А. Ильвовский
«___» _____ 2021 г.

УТВЕРЖДАЮ
Академический руководитель
образовательной программы
«Программная инженерия»,
профессор департамента программной
инженерии, канд. техн. наук

_____ В. В. Шилов
«___» _____ 2021 г.

# ОТЧЕТ О НАУЧНО-ИССЛЕДОВАТЕЛЬСКОЙ РАБОТЕ
# РАЗРАБОТКА ПРОТОТИПА ДЛЯ ГЕНЕРАЦИИ ТЕКСТОВ НА ЕСТЕСТВЕННОМ ЯЗЫКЕ, ОБЛАДАЮЩИХ НАБОРОМ ЗАДАННЫХ СВОЙСТВ

Выполнил
студент группы БПИ184
образовательной программы
09.03.04 «Программная инженерия»

_____ Р.А. Нуртдинов

«___» _____ 2021 г.

Москва 2021

## СПИСОК ИСПОЛНИТЕЛЕЙ

Руководитель НИР, канд. техн. наук _____ Д. А. Ильвовский

Исполнители:

Студент 3-го курса _____ А. Д. Романов
ПИ ФКН НИУ ВШЭ

Студент 3-го курса _____ Р. А. Нуртдинов
ПИ ФКН НИУ ВШЭ

**РЕФЕРАТ**

Отчет 33945 с, 3 рис., 5 табл., 27 источн., 1 прил.

**Ключевые слова:**

*Машинное обучение, кроссвалидация, нейронная сеть, NLP, генерация естественного текста, обучение и построение модели, улучшение модели метрики, ансамблевые методы, GPT-2, ruGPT-3, BERT, Word2Vec.*

В отчете представлены результаты курсовой работы на тему “Разработка прототипа для генерации текстов на естественном языке, обладающих набором заданных свойств“, основанием для НИР является учебный план подготовки бакалавров по направлению 09.03.04 "Программная инженерия" и утвержденная академическим руководителем тема курсового проекта.

**Объект исследования:**

Объектом исследования являются генеративные модели и методы улучшения качества генерируемого текста.

**Цель исследования:**

Разработать прототип для генерации текстов на естественном языке, обладающих набором заданных свойств.

**Методы проведения работы:**

- изучение существующих моделей и библиотек для обработки текстовых данных;
- разработка модели постобработки текста;
- эксперимент и сравнительный анализ.

**Результаты работы:**

Разработан прототип для генерации текстов на основе существующих моделей, позволяющий генерировать текст на естественном языке с набором заданных свойств. Проведено статистическое сравнение нескольких моделей, в том числе с отечественными образцами.

**Область применения результатов:**

Результаты данной работы могут быть использованы для построения русскоязычных генеративных моделей, а также использоваться для решения различных задач в области NLP.

**Рекомендации по внедрению результатов НИР:**

Разработка и реализация программного обеспечения для качественной генерации текста.

**Значимость работы:**

Выводы, сделанные в данной работе, могут использоваться банками и другими финансовыми организациями для улучшения качества текста, генерируемого автоматически.

**Прогнозные предположения о развитии объекта исследования:**

Разработка новых методов обработки текста и комбинация существующих может позволить улучшить качество генерируемого текста.

**СОДЕРЖАНИЕ**

## ТЕРМИНЫ И ОПРЕДЕЛЕНИЯ

В настоящем отчете о НИР применяют следующие термины с соответствующими определениями (таблица 1).

Таблица 1 – Термины и определения

| Термин | Определение |
|---|---|
| Генеративная модель | Модель, самостоятельно производящая текстовые данные |
| Дообучение | Одна из методик улучшения качества генерации текстов, специализированной тематики |
| Естественный язык | Язык, используемый для общения людей |
| Искусственный интеллект | Способность интеллектуальных систем выполнять творческие задачи |
| Кластер слов | Набор слов |
| Корпус | Набор (множество) текстов |
| Лемматизация | Приведение слова к лемме (начальной форме) |
| Машинное обучение | Раздел искусственного интеллекта, изучает методы написания алгоритмов, которые способны обучаться |
| Морфологический анализатор | Набор алгоритмов, занимающихся соотнесением слов и словоформ в лексиконе |
| Обучение модели | Подбор весов числовой функции, описывающей модель, таким образом, чтобы модель выполняла поставленную задачу с максимальной точностью |
| Префикс | Текст, подающийся на вход моделям |
| Таргетированная реклама | Способ рекламировать товар в интернете при помощи совокупности методов и уточнения аудитории для данной рекламы |
| Тематическая модель | Совокупность кластеров и правила построения кластеров |
| Токен | Последовательность символов в документе, которые представляют собой |

| | |
|---|---|
| | семантическую единицу |
| Эмбеддинг | Результат преобразования языковой сущности в числовой вектор |
| Эмодзи | Совокупность идеограмм, используемых при общее на электронных ресурсах |
| API (Application Programming Interface) | Набор функций, которые предоставляет программа внешним пользователям |
| GPT (Generative Pre-trained Transformer) | Поколения алгоритма обработки естественного языка от компании OpenAI |
| GPU (Graphics Processing Unit) | Часть аппаратной платформы, специализирующаяся на графическом рендеринге |
| NLP (Natural Language Processing) | Раздел машинного обучения и математической лингвистики, изучает методы анализа и синтеза естественной речи |

## 1 Введение

Natural Language Processing (далее – NLP) является широко распространенным и быстро развивающимся направлением искусственного интеллекта [1]. Одной из актуальных задач, решаемых NLP, является генерация текстов на естественном языке. Популярность направления выросла с выходом статьи Attention Is All You Need [2], в которой была представлена архитектура Transformer. Массовый интерес появился с выходом моделей Generative Pre-trained Transformer (далее – GPT) [3] [4]. Данные модели предназначены для решения задач на английском языке. Для русского языка аналогичная модель ruGPT-3 была разработана Сбером [5].

Область имеет широкое применение в маркетинге для генерации таргетированной рекламы, а также e-mail ответов для пользователей [6]. Кроме того, текстовые генеративные модели могут использоваться для составления текста новостей и отзывов о товарах.

Однако качество генерируемых текстов далеко от идеала. В приведенном ниже примере использовалась модель ruGPT-3 размера XL. Здесь и далее примеры работы моделей будут приводится в следующем формате: “In: <начало_текста>. Out: <результат_генерации>”, где вместо <начало_текста> подставляется начальное предложение или отрывок, поданный на вход генеративной модели для дополнения (префикс), а вместо <результат_генерации> подставляется дополненный моделью текст.

In: Физики создали квантовую голограмму без прямого наложения двух световых волн.

Out: Физики создали квантовую голограмму без прямого наложения двух световых волн. Вместо этого они использовали взаимосвязь запутанных фотонов, чтобы получить необходимую для построения изображения информацию.http://www/newsrubcapital

Несмотря на грамматическую корректность сгенерированного текста, его содержание сложно назвать осмысленным. Во-первых, заданная тема слабо сохраняется сквозь предложения: в первом предложении речь идет о голограммах, а во втором – уже о фотонах и информации для построения изображения; в третьем предложении приводится некорректная ссылка, не имеющая отношения к теме повествования. Во-вторых, с точки зрения логики предложение 2 не может быть связано с предложением 1, поскольку в первом не содержится отрицаний или потенциальных объектов замещения, после которых может использоваться предлог “вместо”.

Другой пример логически некорректного текста представлен ниже (вновь используется ruGPT-3 размера XL).

In: По оценке Банка России, восстановление экономической активности в конце 2020 года продолжилось. На это оказали влияние несколько факторов.

Out: По оценке Банка России, восстановление экономической активности в конце 2020 года продолжилось. На это оказали влияние несколько факторов. В частности рост промышленного производства составил 0%. Кроме того улучшилось состояние внешней торговли (0% по сравнению с 2017 годом) и снизились цены на нефть...

Здесь вместо перечисления факторов модель раскрывает, в чем заключается восстановление экономической активности. А также говорит об улучшении некоторых показателей, в то же время приводя разницу в 0%.

Таким образом, модель проигрывает по нескольким характеристикам текста.

В данном исследовании оцениваются результаты генерации нескольких моделей и рассматриваются методы улучшения качества генерируемого текста по ряду показателей без изменения архитектуры модели.

## 2 Направление исследования

Направлением данного исследования является раздел искусственного интеллекта NLP, а именно генеративные модели. Данная область была выбрана из-за большого количества открытых проблем, которые описаны во многих статьях [7]. Задача генерации текста привлекла своей актуальностью, наличием большого количества открытых моделей (в частности, моделей, обученных на корпусе русского языка от Сбера (ruGPT–3)), а также своей популярностью в последний год.

Написание и обучение собственной модели – трудоемкий и ресурсоемкий процесс, который под силу немногим организациям. Поэтому основной задачей исследования ставилось улучшение качества генерируемого текста готовой модели без изменения ее архитектуры. Улучшение осуществимо многими подходами, а результат, в основном, измеряется эмпирически, после чего делаются выводы о применимости тех или иных методов. Методы улучшения, предложенные данным исследованием, приведены в п.4.

## 3 Анализ существующих моделей и библиотек

В рамках исследования был проведен анализ существующий генеративных моделей библиотек для работы с текстами. Описание моделей GPT, T5, BERT было частью общей исследовательской работы, однако оно выполнялось не мной [8] [9] [10] [11] [12] [13] [14].

### 3.1 Word2vec

Библиотека Word2vec использовалась из реализации в библиотеке Gensim [15]. Ее основной задачей является преобразование слов в векторное представление (эмбеддинги) на основе их взаимного расположения в обучающем корпусе. В текущем исследовании библиотека использовалась для подсчета косинусного расстояния между словами и впоследствии поиска слов наиболее схожих с данным. Модель выбрана благодаря всеобщему признанию и быстроте работы в данной реализации.

### 3.2 BigARTM (LDA)

BigARTM [16] – это библиотека с открытым исходным кодом, написанная на языке C++. Позволяет выполнять тематическое моделирование больших текстовых данных. Использование данной библиотеки сводилось к составлению тематической модели слов, которые встречались в корпусе. Для построения модели был использован алгоритм LDA [17]. Данная библиотека была выбрана из-за заявленной эффективной потоковой параллельной реализации.

### 3.3 UDpipe

UDpipe [18] – библиотека для сопоставления словам тегов частей речи (определение частей речи, так называемый POS-tagging), лемматизации и синтаксического анализа. Первые две возможности активно использовались в настоящем исследовании. Поводом для выбора библиотеки стала ее хорошая адаптированность к русскому языку.

### 3.4 pymorphy

Морфологический анализатор pymorphy [19] использовался для приведения слова к нормальной форме, а также постановки слова в нужное склонение. Он тоже неплохо адаптирован для русского языка и достаточно прост в использовании.

## 4 Методы улучшения качества генерируемого текста

В связи с ранее упомянутыми проблемами, BERT и T5 не могли использоваться как генеративные модели в рамках исследования. Поэтому было решено сфокусироваться на разных видах GPT.

Оригинальная GPT-3 имеет платный API [20], поэтому ее использование не представлялось возможным. Исходя из этого в исследовании для улучшения рассматривались исходная английская GPT-2 и русская ruGPT-3. Обе модели были взяты размера S.

### 4.1 Англоязычная модель с переводчиком

Исходная GPT-2 предобучена на английском корпусе и не имеет русскоязычных аналогов. Она генерирует тексты на английском языке и крайне плохо справляется с русским.

В рамках исследования к модели был подключен переводчик: идея заключалась в том, чтобы переводить префикс на английский, подавать на вход генеративной англоязычной GPT-2, а затем переводить результат генерации на русский. Описанная схема приведена на рисунке 1.

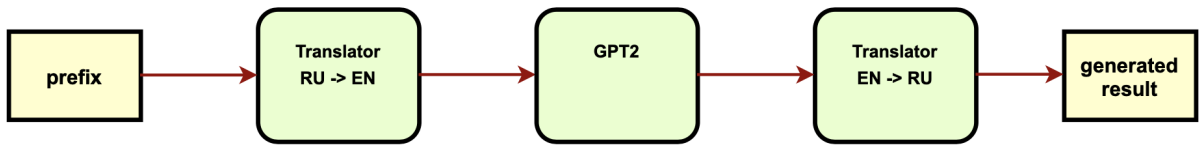


Рисунок 1 – Модель с двусторонним переводчиком

Для реализации использовался API Google Cloud Translation [21]. Достоинством данного API являлась возможность бесплатного пользования практически без ограничений. Однако качество перевода с английского на русский было неудовлетворительным: имели место непереведенные фразы, и большое число грамматических ошибок.

В связи с этим был также настроен Yandex Translate API на платформе Yandex Cloud [22]. Взаимодействие с ним было выстроено с помощью HTTP API. Качество его перевода было очень высоким, практически не возникали ошибки в склонениях слов. В связи с этим было решено использовать генерацию GPT2 именно в связке с Yandex Translate.

### 4.2 Дообучение модели

Исходные модели GPT-2 и ruGPT-3 предобучены на текстах самой разной тематики. Для улучшения соответствия генерируемого текста определенной тематике, в частности, модели предусматривают API для дообучения. На вход подается простой неразмеченный текст, а после достаточно продолжительного процесса вычислений получившиеся состояния модели сохраняются в так называемые чекпоинты (checkpoints), которые в дальнейшем могут быть загружены моделью и впоследствии использоваться для генерации. Чем больше объем текста, подаваемый для дообучения, тем большее влияние дообучение оказывает на модель. Подробнее о данных, использовавшихся для дообучения см. п.5.

Отдельно стоит заметить, что в случае с GPT-2 те же данные предварительно были переведены на английский язык переводчиком (см. п.4.1), и только затем поданы для дообучения.

Исходный код для дообучения и последующей генерации моделями GPT-2[1] и ruGPT-3[2] представлен в среде Google Colab.

### 4.3 Замена слов на синонимы из тематической модели

Замены слов на синонимы нередко используются для искусственного увеличения корпуса в NLP. В рамках исследования была собрана модель замен, которая, используя ряд библиотек, заменяет слова в результате генерации на синонимичные из соответствующего тематического словаря (далее – модель замен). Это позволяет добиться использования более специфичной, соответствующей выбранной тематике терминологии в генерируемом тексте без потери его основной мысли. Далее приводится описание полученной модели.

Первоочередной задачей был сбор словаря терминологии определенной тематики. Изначально предполагалось тематическое моделирование при помощи библиотеки BigARTM. Однако результат ее работы оказался неудовлетворительным: внутри каждого кластера находились слова самой разной тематики, что не позволяло выбрать кластер, где все слова относились бы к какой-то одной конкретной тематике. Это впоследствии не позволяло составить упомянутый тематический словарь. Лемматизация и фильтрация по частям речи (оставили лишь существительные и прилагательные) также не позволили добиться желаемого результата. Тогда было решено собрать словарь вручную. Описание его сбора приведено в п.5.

[1] https://colab.research.google.com/drive/1nhrtXwNU1vFsVPn25zf21CbqUkaU7sn6
[2] https://colab.research.google.com/drive/13xPwb-UlYgblOeA8eFr_sK6MAOlPRNWi

Следующим шагом был алгоритм подбора синонимичных слов. Здесь использовался word2vec, реализованный в библиотеке Gensim. Однако предобученный он не содержал многие слова из полученного словаря специфичных терминов. В связи с этим возникла потребность в дообучении. При помощи библиотеки UDpipe были получены теги частей речи и начальные формы для всех слов собранного корпуса [23], а затем поданы для дообучения word2vec [24]. Полученные состояния были сохранены в чекпоинты аналогично дообученным моделям GPT. word2vec теперь мог генерировать для заданного слова синонимы из специфичного словаря. Дообучение word2vec также представлено в Google Colab[3].

Далее пошагово рассматривается генерация и обработка текста при вызове.

1) Поданный на вход префикс передается дообученной (см п.4.2) генеративной модели GPT (GPT-2 с переводчиком (см. п.4.1) или ruGPT-3).

2) Результат генерации передается в UDPipe, где с помощью тегов частей речи всем словам сопоставляются их части речи, а затем фильтруются лишь прилагательные и существительные (так как именно они чаще всего являются частью специфичной терминологии) – потенциальные кандидаты на замену.

3) Кандидаты на замену передаются в дообученный word2vec, где к каждому слову подбирается набор синонимов с учетом части речи и подсчитывается косинусное расстояние до изначального слова. В данном модуле все слова используются в начальной форме.

4) Синонимы фильтруются по наличию в банковском словаре (его формирование описано ранее в этом пункте), таким образом, остаются лишь синонимы, относящиеся к специфичной тематике.

5) Тематические синонимы склоняются с помощью pymorphy, принимая форму заменяемого слова.

6) Слово, подлежащее замене, поочередно заменяется в выводе генеративной модели на очередной склоненный тематический синоним в порядке убывания косинусного расстояния до заменяемого слова (рисунок 2). При этом отсеиваются синонимы, имеющие слишком малое (то есть значительно далекие по значению) косинусное расстояние. Минимальный порог задается вручную.

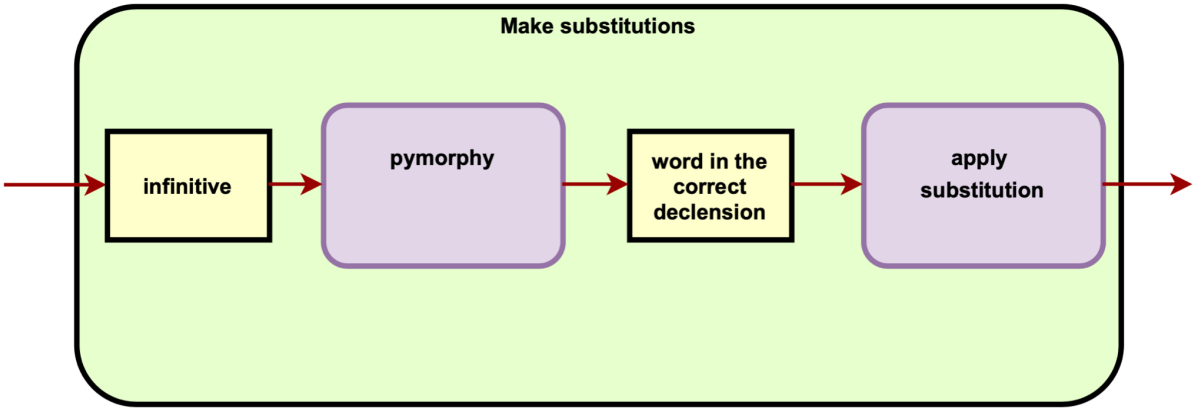


Рисунок 2 – Модуль применения замен в модели замен

[3] https://colab.research.google.com/drive/1kDwHubiey3t5YeS0dymH1aB4ptwEvKB1

7) Между результатом генерации и результатом с выполненной заменой подсчитывается косинусное расстояние при помощи MultiBERT. Таким образом, подбирается наиболее подходящая замена, то есть та, при которой косинусное расстояние между двумя результатами наиболее близко к 1. При этом, аналогично предыдущему пункту, отсеиваются замены, имеющие слишком малое косинусное расстояние. Минимальный порог так же задается вручную.

8) Результат лучшей замены сохраняется и становится выводом модели замен.

Целиком модель замен схематично изображена на рисунке 3, а также опубликована[4] в Google Colab.

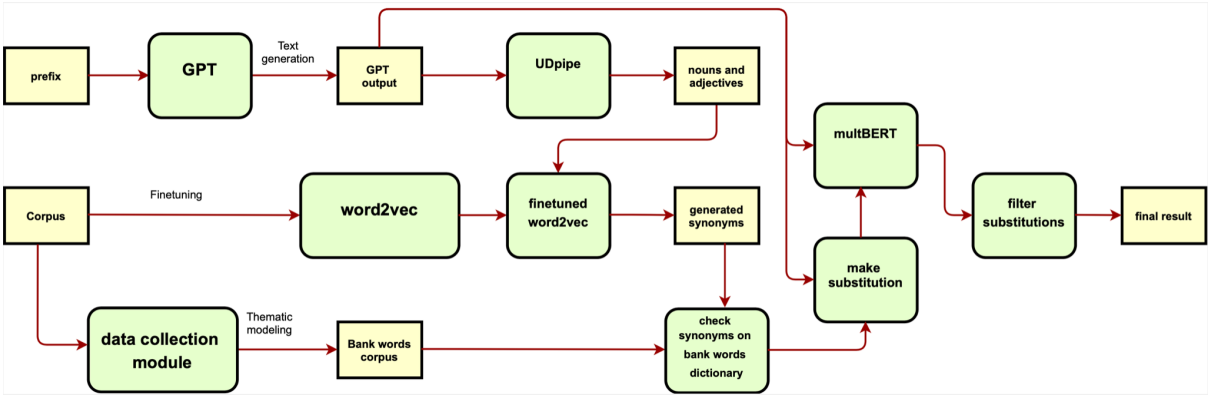


Рисунок 3 – Схема модели замен

## 5 Данные

В рамках подготовки данных для эксперимента были собраны словарь банковских терминов и текст длиной 1790 символов (13901 статья). Данная сфера была выбрана в связи со своей специфичностью, а также компетентности исполнителей исследования в ней. Сбор этих данных был частью общей исследовательской работы, однако выполнялся не мною [25].

## 6 Экспериментальные исследования

Для демонстрации прикладного аспекта исследования был проведен эксперимент.

### 6.1 Описание методологии проведения эксперимента

Экспериментальная часть работы состоит из двух частей.

Первая часть заключается в выборе моделей, сборе тестового корпуса префиксов и генерации некоторого количества текстов по заданным образцам банковской тематики.

[4] https://colab.research.google.com/drive/1zRGrZNqbJzW5mKcVUfyH8ylxXiGJ-TbF

Для эксперимента использовались следующие варианты моделей: GPT-2 размера S, ruGPT-3 размера S, дообученная GPT-2 размера S с алгоритмом замены синонимов без проверки замен на принадлежность банковскому словарю, дообученная GPT-2 размера S с алгоритмом замены на схожие слова банковской тематики. В последних двух был выключен модуль MultiBERT в связи с его сильным замедлением работы алгоритма. Из двух GPT моделей для алгоритма замен была выбрана GPT-2, так как по результатам тестов показала себя лучше.

Текстовый корпус префиксов для генерации состоял из 30 текстов и был собран с тематических сайтов: Банки.ру и ЦБРФ[5].

Вторая часть эксперимента заключается определении статистических гипотез, оценке сгенерированных текстов по нескольким критериям и статистической проверке гипотез при помощи различных критериев.

### 6.2 Оценивание качества текстов

Для оценки качества генерируемого текста были выбраны следующие характеристики: грамматическая корректность, соответствие теме, логическая связность. Данные характеристики были выбраны, поскольку именно по ним генеративные модели чаще всего проигрывают.

Ввиду отсутствия программного обеспечения для автоматической проверки соответствия текста теме и логики повествования, оценка каждого текста производилась вручную по всем трем параметрам по шкале от 0 до 5. Для этого были разработаны соответветствующие критерии, приведенные в таблицах 2, 3, 4.

Таблица 2 – Критерии оценивания текста на грамматическую корректность

| Оценка | Критерий |
|---|---|
| 0 | Текст содержит более 15 грамматических ошибок (в склонениях слов, орфографии) |
| 1 | Текст содержит 10-15 грамматических ошибок |
| 2 | Текст содержит 7-9 грамматических ошибок |
| 3 | Текст содержит 4-6 грамматических ошибок |
| 4 | Текст содержит от 2-3 грамматические ошибки |
| 5 | Текст содержит 1 или менее грамматических ошибок |

Таблица 3 – Критерии оценивания текста на соответствие теме

[5] https://www.cbr.ru/

| Оценка | Критерий |
|---|---|
| 0 | Сгенерированная часть текста крайне далека по тематике от заданного префикса или состоит из неразборчивых символов |
| 1 | Сгенерированная часть текста далека по тематике от заданного префикса |
| 2 | Сгенерированная часть текста далека по тематике от заданного префикса, однако содержит некоторые термины из данной тематики |
| 3 | Как минимум половина сгенерированной части текста соответствует заданному префиксу по тематике |
| 4 | Сгенерированная часть текста в целом соответствует по тематике заданному префиксу, однако в допускаются незначительные отхождения от темы |
| 5 | Сгенерированная часть текста полностью соответствует по тематике заданному префиксу |

Таблица 4 – Критерии оценивания текста на логическую связность

| Оценка | Критерий |
|---|---|
| 0 | Текст полностью противоречивый и/или представляет из себя набор несвязных фраз или слов |
| 1 | Текст содержит логические противоречия и/или практически лишен связности |
| 2 | Текст не характеризуется связностью, присутствуют “вырванные” фразы, резкие переходы без слов-связок, логические противоречия практически отсутствуют |
| 3 | Текст в целом связный, логические противоречия отсутствуют, однако допускается наличие небольшого количества несвязных фраз |
| 4 | Текст характеризуется связностью, логические противоречия отсутствуют, однако допускаются небольшие отклонения в логике повествования |
| 5 | Текст характеризуется связностью, целостностью, ясностью, логические противоречия отсутствуют, в тексте присутствуют слова-связки, не нарушена логика повествования |

Результаты оценивания текстов тестового корпуса по данным критериям находятся в приложении А. В таблице 5 приведены средние значения полученных оценок.

Таблица 5 – Средние оценки текстов, сгенерированных в процессе эксперимента[6]

| грамматическая корректность | | | | соответствие темы | | | | логическая связность | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| gpt3 | gpt2 | зам. | б.зам. | gpt3 | gpt2 | зам. | б.зам. | gpt3 | gpt2 | зам. | б.зам. |
| 4,6 | 3,5 | 3,0 | 4,4 | 2,9 | 2,1 | 3,1 | 3,7 | 3,4 | 3,1 | 3,3 | 4,0 |

### 6.3 Методы расчета

Для проверки статистических гипотез был выбран критерий Уилкоксона [26] по причине относительно небольшого размера выборки текстов (30 образцов) и определении количественных критериев оценки текста в соответствии с п.6.2.

### 6.4 Основание для проведения экспериментальных работ

Основанием для проведения экспериментальных работ является отсутствие программных решений, позволяющих оценить качество генерируемых текстов в требуемом формате и отсутствие возможности теоретически принять или опровергнуть сформулированные гипотезы.

## 7 Обобщение и оценка результатов исследования

### 7.1 Оценка результатов исследования

Результаты исследования были подтверждены проверкой статистических гипотез [27]. Они показали, что модель замен (с банковским словарем) имеет преимущества по показателям логической связности и соответствию теме, однако, согласно расчетам, оснований принять гипотезу об улучшении грамматической корректности не было. Проверка статистических гипотез была частью общей исследовательской работы, однако оно выполнялось не мной.

### 7.2 Оценка достоверности полученных результатов

[6] gpt3 – оценка текстов, сгенерированных моделью ruGPT-3 размера S
gpt2 – оценка текстов, сгенерированных моделью GPT-2 размера S
зам. – дообученная GPT-2 размера S с алгоритмом замены синонимов без проверки замен на принадлежность банковскому словарю
б.зам. – дообученная GPT-2 размера S с алгоритмом замены на схожие слова банковской тематики

Результаты можно считать достоверными поскольку оценка текстов проводилась несколькими людьми независимо друг от друга, данные для генерации текстов выбирались случайным образом с тематических ресурсов, для проверки гипотез применялся непараметрический статистический тест.

### 7.3 Дополнительные исследования

Развитием данной научно-исследовательской работы может стать новая серия экспериментов с применением более “тяжелых” моделей, обученных под другие специфичные тематики на новых корпусах текстов. Другим вариантом развития данной работы может стать увеличение количества людей, проводящих оценку текстов и изменение критериев оценки с целью получения более достоверных результатов. Также рассматривается вариант написания программного обеспечения для автоматической оценки сгенерированного текста по заданным параметрам с целью ускорить и автоматизировать процесс проверки текста, что позволит под новым углом взглянуть на задачи верификации текстов.

В настоящий момент выбранная в работе область машинного обучение переживает бурное развитие, в связи с чем количество открытых проблем увеличивается с каждым днем, что позволяет надеяться на большой скачок в этой области в ближайшее время.

### 7.4 Отрицательные результаты

Отрицательными результатами можно считать результаты проверок первых двух гипотез, которые показали, что новая модель в рамках поставленного эксперимента не имеет статистически значимых улучшений в сравнении с уже имеющимися моделями.

На протяжении выполнения исследовательской работы также были замечены следующие недостатки современных генеративных моделей:

– высокие системные требования генеративных моделей и сопутствующих им библиотек как к программной части, так и к аппаратной, которые не позволили запустить некоторые модели на локальном компьютере;

– точечные отклонения от темы повествования в сгенерированном тексте лучших моделей и значительное отхождение от темы повествования в моделях предыдущего поколения;

– проблемы с проверкой фактов в генерируемом тексте (информация подаваемая как фактическая зачастую абсурдна и не несет никакой полезной смысловой нагрузки).

## ЗАКЛЮЧЕНИЕ

Результаты НИР показали, что данная задача NLP имеет большое количество открытых проблем, которые предстоит решить в будущих исследованиях.

Одной из основных задач данной НИР являлось создание прототипа для генерации текста на естественном языке, обладающих набором заданных свойств с улучшенным качеством генерируемого текста. Пример его работы приведен ниже.

In: Рубль завершил основную валютную сессию четверга укреплением к доллару и евро

Out: Рубль завершил основную валютную сессию в четверг укреплением по отношению к доллару и евро, передает ТАСС. Нефтяные котировки доллара и евро потеряли за неделю около 40%, отмечает старший аналитик Информационно-аналитического центра Альпари в РБА Анастасия Кролова. В настоящее время на фоне заседания ОПЕК у рынка нет особого аппетита к дополнительным санкциям против стран, не подписавших Соглашение по ОПЕК +. Еще 10 участников плана включают восемь стран - Россию, Казахстан, Белоруссию, Армению, Армению. Как пояснила премьер-министр РФ, для борьбы с коррупцией и мошенничеством Россия будет говорить о расширении сотрудничества и решении оперативных задач в этой сфере, и мы будем говорить об участии в этом межгосударственных организаций.

Таким образом, собранный прототип уверенно показал себя в сравнении с существующими решениями: статистически значимо улучшилась логическая связность и возросло соответствие теме в результатах генерации – это два ключевых параметра, по которым обычно проигрывают генеративные модели. Тем не менее, как показала проверка первых двух гипотез, грамматическая корректность не возросла, однако это и не являлось первоочередной целью данного исследования.

Собранный корпус банковских слов можно использовать для дообучения других тематических моделей, а соответствующие дообученные модели GPT-2, ruGPT-3 размера S и word2vec – для продолжения исследований генеративных моделей и решения других задач NLP, особенно на русском языке. Приведенный в работе алгоритм замен можно применять для других естественных языков.

Результаты работы опубликованы в публичном репозитории[7].

[7] https://github.com/nitrochange/finetuning-ruGPT3

## СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ


1. Natural Language Processing Basics for Text. [Electronic resource] / Habr [Official website]. URL: https://habr.com/ru/company/Voximplant/blog/446738/ (accessed: 01.03.2021)
2. Polosukhin I., Kaiser L., Parmar N. Attention Is All You Need / Arxiv [Official website]. URL: https://arxiv.org/abs/1706.03762 (accessed: 21.03.2021)
3. Radford A., Luan D.,Amodei D., Sutskever I., Language Models are Unsupervised Multitask Learners [Electronic resource] / Arxiv [Official website]. URL: https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf (accessed: 20.03.2021)
4. Sutskever I., Ramesh A., Dhariwal P., Neelakantan A., Language Models are Few-Shot Learners. [Electronic resource] / Arxiv [Official website]. URL: https://arxiv.org/abs/2005.14165 (accessed: 15.03.2021)
5. ruGPT-3: генеративная русскоязычная нейросетевая модель. [Электронный ресурс] / Sbercloud [Официальный ресурс]. URL: https://sbercloud.ru/ru/warp/gpt-3 (дата обращения: 15.03.2021)
6. Natural Language Generation and Its Business Applications. [Electronic resource] / Skimai. URL: https://skimai.com/natural-language-generation-business-applications/ (accessed: 20.12.2020)
7. Ruder S., The 4 Biggest Open Problems in NLP. [Electronic resource]. / Ruder [Official website]. URL: https://ruder.io/4-biggest-open-problems-in-nlp/ (accessed: 23.03.2021)
8. BERT. [Electronic resource]. / Github [Official website]. URL: https://github.com/google-research/bert/blob/master/multilingual.md (accessed: 03.03.2021)
9. Understanding searches better than ever before. [Electronic resource]. / Google Blog [Official website]. URL: https://blog.google/products/search/search-language-understanding-bert/ (accessed: 20.01.2021)
10. Rogers A., Kovaleva O., Rumshisky A., A Primer in BERTology: What We Know About How BERT Works. [Electronic resource]. / Arxiv [Official website]. URL: https://arxiv.org/pdf/2002.12327.pdf (accessed: 25.03.2021)
11. Cosine similarity. [Electronic resource]. / Wikipedia [Official website]. URL: https://en.wikipedia.org/wiki/Cosine_similarity (accessed: 09.03.2021)
12. Exploring Transfer Learning with T5: the Text-To-Text Transfer Transformer. [Electronic resource]. / AI Google blog [Official website]. URL: https://ai.googleblog.com/2020/02/exploring-transfer-learning-with-t5.html (accessed: 06.12.2020)
13. Google Colaboratory. [Electronic resource]. / AI Google Research [Official website]. URL: https://colab.research.google.com/ (accessed: 09.03.2021)
14. Better Language Models and Their Implications. [Electronic resource]. / OpenAI [Official website]. URL: https://openai.com/blog/better-language-models/ (accessed: 05.04.2021)

15. Gensim. [Electronic resource]. / Gensim [Official website]. URL: https://radimrehurek.com/gensim/ (accessed: 28.03.2021)
16. BigARTM. [Electronic resource]. / AI Google Research [Official website]. URL: https://bigartm.readthedocs.io/en/stable/ (accessed: 24.03.2021)
17. Латентное размещение Дирихле. [Электронный ресурс]. / Латентное размещение Дирихле [Официальный ресурс]. URL: https://ru.wikipedia.org/wiki/Латентное_размещение_Дирихле (дата обращения: 26.03.2021)
18. UDPipe. [Electronic resource]. / UDPipe [Official website]. URL: https://ufal.mff.cuni.cz/udpipe (accessed: 10.04.2021)
19. Морфологический анализатор pymorphy2.[Электронный ресурс]. / pymorphy2 [Официальный ресурс]. URL: https://pymorphy2.readthedocs.io/en/stable/ (дата обращения: 17.03.2021)
20. Is OpenAI's GPT-3 API Beta Pricing Too Rich for Researchers? [Electronic resource] / Syncedreview [Official website]. URL: https://syncedreview.com/2020/09/04/is-openais-gpt-3-api-beta-pricing-too-rich-for-researchers/ (accessed: 20.02.2021)
21. Google Cloud Translation. [Electronic resource] / Google translator API [Official website]. URL: https://cloud.google.com/translate/?hl=ru (accessed: 10.12.2020)
22. Yandex Cloud Translation. [Electronic resource] / Yandex translator API [Official website]. URL: https://cloud.yandex.ru/services/translate (accessed: 01.04.2021)
23. Обучаем Word2vec: практикум по созданию векторных моделей языка. [Электронный ресурс]. / Sysblock [Официальный ресурс]. URL: https://sysblok.ru/knowhow/obuchaem-word2vec-praktikum-po-sozdaniju-vektornyh-modelej-jazyka/ (дата обращения: 19.01.2021)
24. Tatman R., Fine tuning word2vec. [Electronic resource]. / Kaggle [Official website] URL: https://www.kaggle.com/rtatman/fine-tuning-word2vec (accessed: 11.02.2021)
25. Банковский словарь. [Электронный ресурс]. / Банк справка [Официальный ресурс]. URL: https://bankspravka.ru/bankovskiy-slovar/bankovskiy-slovar.html#o (дата обращения: 15.02.2021)
26. W критерий Уилкоксона. [Электронный ресурс]. / W критерий Уилкоксона [Официальный ресурс]. URL: http://statistica.ru/local-portals/medicine/w-kriteriy-uilkoksona/ (дата обращения: 10.04.2021)
27. Критические значения критерия Уилкоксона. [Электронный ресурс]. / Критические значения критерия Уилкоксона [Официальный ресурс]. URL: https://gymnasium42.ru/stat/Book/Data/page_7.htm (дата обращения: 12.04.2021)

## ПРИЛОЖЕНИЕ А

## РЕЗУЛЬТАТЫ ОЦЕНИВАНИЯ ТЕКСТОВ ТЕСТОВОГО КОРПУСА

Полные тексты (префиксы и результаты генерации) представлены в Google-таблице[8]. Результаты оценивания продублированы в настоящем приложении в таблице А.1.

Таблица А.1 – Результаты оценивания текстов тестового корпуса[9]

| Номер префикса | Номер результата генерации | Грамматическая корректность | Соответствие темы | Логическая связность |
|---|---|---|---|---|
| | | | | |
| 1 | 1 | 4 | 3 | 4 |
| | 2 | 5 | 3 | 2 |
| | 3 | 4 | 4 | 3 |
| | 4 | 5 | 4 | 4 |
| | | | | |
| 2 | 1 | 5 | 3 | 3 |
| | 2 | 4 | 0 | 1 |
| | 3 | 3 | 2 | 2 |
| | 4 | 5 | 4 | 3 |
| | | | | |
| 3 | 1 | 4 | 3 | 3 |
| | 2 | 3 | 4 | 3 |
| | 3 | 2 | 3 | 3 |
| | 4 | 4 | 3 | 4 |
| | | | | |
| 4 | 1 | 5 | 4 | 4 |

[8] https://docs.google.com/spreadsheets/d/1AX6crxt_F2RGDyF7fkgDkjBrkyB1knqw036PBManlk0

[9] Номера результатов генерации:
1 – оценка текстов, сгенерированных моделью ruGPT-3 размера S
2 – оценка текстов, сгенерированных моделью GPT-2 размера S
3. – дообученная GPT-2 размера S с алгоритмом замены синонимов без проверки замен на принадлежность банковскому словарю
4 – дообученная GPT-2 размера S с алгоритмом замены на схожие слова банковской тематики

| | | | | |
|---|---|---|---|---|
| | 2 | 2 | 2 | 2 |
| | 3 | 4 | 3 | 1 |
| | 4 | 4 | 5 | 4 |
| | | | | |
| 5 | 1 | 4 | 1 | 2 |
| | 2 | 4 | 3 | 3 |
| | 3 | 1 | 2 | 3 |
| | 4 | 4 | 3 | 3 |
| | | | | |
| 6 | 1 | 5 | 5 | 5 |
| | 2 | 4 | 2 | 2 |
| | 3 | 3 | 3 | 4 |
| | 4 | 4 | 3 | 4 |
| | | | | |
| 7 | 1 | 5 | 3 | 5 |
| | 2 | 3 | 1 | 2 |
| | 3 | 2 | 3 | 3 |
| | 4 | 5 | 3 | 3 |
| | | | | |
| 8 | 1 | 5 | 4 | 4 |
| | 2 | 3 | 5 | 4 |
| | 3 | 2 | 3 | 4 |
| | 4 | 5 | 4 | 3 |
| | | | | |
| 9 | 1 | 4 | 2 | 4 |
| | 2 | 4 | 4 | 3 |
| | 3 | 1 | 3 | 4 |
| | 4 | 5 | 4 | 4 |
| | | | | |
| 10 | 1 | 4 | 3 | 2 |
| | 2 | 3 | 2 | 3 |
| | 3 | 3 | 1 | 2 |
| | 4 | 4 | 4 | 3 |

| | | | | |
|---|---|---|---|---|
| 11 | 1 | 4 | 3 | 3 |
| | 2 | 5 | 4 | 3 |
| | 3 | 4 | 3 | 2 |
| | 4 | 5 | 4 | 4 |
| | | | | |
| 12 | 1 | 4 | 2 | 2 |
| | 2 | 4 | 1 | 3 |
| | 3 | 5 | 2 | 4 |
| | 4 | 4 | 4 | 4 |
| | | | | |
| 13 | 1 | 4 | 4 | 3 |
| | 2 | 2 | 3 | 5 |
| | 3 | 3 | 4 | 4 |
| | 4 | 4 | 3 | 5 |
| | | | | |
| 14 | 1 | 5 | 3 | 4 |
| | 2 | 3 | 2 | 4 |
| | 3 | 4 | 1 | 3 |
| | 4 | 5 | 3 | 4 |
| | | | | |
| 15 | 1 | 5 | 4 | 4 |
| | 2 | 4 | 3 | 4 |
| | 3 | 2 | 3 | 3 |
| | 4 | 5 | 3 | 5 |
| | | | | |
| 16 | 1 | 4 | 2 | 2 |
| | 2 | 2 | 2 | 5 |
| | 3 | 3 | 3 | 5 |
| | 4 | 4 | 3 | 5 |
| | | | | |
| 17 | 1 | 5 | 2 | 4 |
| | 2 | 5 | 1 | 4 |

| | | | | |
|---|---|---|---|---|
| | 3 | 3 | 2 | 2 |
| | 4 | 3 | 4 | 4 |
| | | | | |
| 18 | 1 | 4 | 3 | 4 |
| | 2 | 5 | 3 | 4 |
| | 3 | 3 | 3 | 2 |
| | 4 | 5 | 4 | 5 |
| | | | | |
| 19 | 1 | 5 | 3 | 3 |
| | 2 | 4 | 2 | 3 |
| | 3 | 3 | 4 | 5 |
| | 4 | 3 | 3 | 4 |
| | | | | |
| 20 | 1 | 5 | 1 | 3 |
| | 2 | 4 | 3 | 5 |
| | 3 | 3 | 4 | 4 |
| | 4 | 5 | 3 | 4 |
| | | | | |
| 21 | 1 | 5 | 4 | 5 |
| | 2 | 3 | 3 | 4 |
| | 3 | 4 | 3 | 4 |
| | 4 | 4 | 4 | 5 |
| | | | | |
| 22 | 1 | 5 | 3 | 4 |
| | 2 | 4 | 2 | 2 |
| | 3 | 4 | 3 | 3 |
| | 4 | 5 | 4 | 4 |
| | | | | |
| 23 | 1 | 5 | 3 | 3 |
| | 2 | 3 | 1 | 2 |
| | 3 | 3 | 5 | 4 |
| | 4 | 4 | 5 | 4 |
| | | | | |

| | | | | |
|---|---|---|---|---|
| 24 | 1 | 4 | 1 | 5 |
| | 2 | 3 | 2 | 5 |
| | 3 | 4 | 3 | 4 |
| | 4 | 5 | 2 | 5 |
| | | | | |
| 25 | 1 | 5 | 4 | 4 |
| | 2 | 2 | 0 | 3 |
| | 3 | 3 | 3 | 5 |
| | 4 | 4 | 3 | 4 |
| | | | | |
| 26 | 1 | 5 | 3 | 3 |
| | 2 | 4 | 0 | 3 |
| | 3 | 3 | 3 | 3 |
| | 4 | 4 | 5 | 4 |
| | | | | |
| 27 | 1 | 5 | 4 | 3 |
| | 2 | 3 | 1 | 3 |
| | 3 | 3 | 4 | 4 |
| | 4 | 5 | 3 | 4 |
| | | | | |
| 28 | 1 | 5 | 3 | 2 |
| | 2 | 3 | 0 | 1 |
| | 3 | 1 | 4 | 3 |
| | 4 | 3 | 4 | 4 |
| | | | | |
| 29 | 1 | 5 | 1 | 2 |
| | 2 | 3 | 1 | 3 |
| | 3 | 3 | 5 | 4 |
| | 4 | 4 | 5 | 3 |
| | | | | |
| 30 | 1 | 5 | 2 | 3 |
| | 2 | 5 | 3 | 2 |
| | 3 | 4 | 5 | 3 |

| | | | | |
|---|---|---|---|---|
| | 4 | 5 | 4 | 5 |