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»

_____ А.Д. Романов

«___» _____ 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 34413 с, 3 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.

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	Часть искусственного интеллекта, которая занимается разработкой моделей, способных выполнять творческие действия
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)	Пересечение машинного обучения и математической лингвистики, направленное на изучение методов анализа и синтеза естественного языка

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.

3.1 BERT

BERT [8] – модель, представленная компанией Google в 2018 году, что характеризует данную модель актуальностью. BERT был обучен под большое количество языков, и его публикация считается весьма важной вехой в решении задач NLP. Примечательно, что компания Google внедрила BERT в алгоритм поиска [9]. Сама модель, в силу своей новизны, еще не до конца изучена [10]. Данная модель может решать целый комплекс NLP задач: анализ текста, выявление спама, режим вопросно-ответной системы, суммаризация текста и так далее. В то же время она не предоставляет API для генерации текстов, что помешало ее использованию в качестве модели генерации. В настоящем исследовании использовалась реализация данной модели MultiBERT в библиотеке transformers для расчета косинусного расстояния [11] между предложениями.

3.2 T5

T5 [12] – это также NLP-модель от компании Google. Она является генеративной, однако использовать ее в настоящем исследовании не представилось возможным из-за возникших ошибок при работе с API.

3.3 GPT

Чуть подробнее рассмотрим семейство моделей GPT, созданных и поддерживаемых компанией OpenAI. Данная модель позиционируется как лучшая NLP-модель на момент написания статьи и позволяющая решать любые задачи этой области на английском языке. Для русского языка данную модель собрала, дообучила и выложила в открытый доступ компания Сбер осенью 2020 года.

Семейство GPT на момент написании статьи имеет 3 поколения (GPT-1, GPT-2, GPT-3). Каждое поколение включает в себя несколько различных моделей. Их отличие заключается в количестве параметров обучения. С увеличением числа параметров в модели улучшается ее результат генерации.

Настоящее исследование проводилось в интерактивной среде выполнения Google Colab [13] (далее – Google Colab) по нескольким причинам: данная среда предоставляет функционал для совместной работы, и в ней можно выполнять код моделей с использованием графического ускорителя GPU. Данная возможность является критичной, в силу того, что большинство описанных моделей и библиотек, которые являются важными структурными элементами нашего исследования, не запускаются или теряют в производительности в средах выполнения без графического ускорителя.

В настоящем исследовании использовались “маленькие” модели GPT размера S, так как запуск моделей большего размера в Google Colab не представлялся возможным в силу нехватки памяти ОЗУ (на бесплатной версии выделяется около 12 ГБ такой памяти).

Использовавшаяся ruGPT-3 размера S (125 млн параметров), исходя из информации, выложенной компанией Сбер, была обучена на 80 миллионах токенов в 3 эпохи. Обучение заняло около недели и выполнялось на 32 графических процессорах. Далее приведен демонстрационный пример работы модели в режиме генерации текста:

In: Банки активно используют это волшебное свойство безналичного рубля на благо экономики.

Out: Банки активно используют это волшебное свойство безналичного рубля на благо экономики. Этот платный эффект может иметь радикальные последствия. Сосредоточив внимание на факторах, влияющих на долю рынка, эти санкции могут не иметь большого значения, но вы можете обнаружить более глубокие последствия для фонда спасения, чем проблемы реструктуризации банков. Как этот опыт распространяется на другие отрасли? Но, мягко говоря, Закон о реформировании активов, определяемый регулирующими органами как "поддержка создания рутинной банковской системы", вероятно, увеличивает риск банкротства. В США ежегодно теряется до 180 000 дел о корпоративной собственности - по сравнению с 4000 в 2008 году. Многие из этих требований основаны на недействительности титула, собственности, ипотеки и корпоративных прав, и могут быть удовлетворены "безналичным" рублем всего за 20 000 долларов на ваш дом с надеждой на ежегодное федеральное взыскание в США, но с более чем десятикратной вероятностью наступления срока погашения.

GPT-2 размера S (124 млн параметров) была выпущена компанией OpenAI в феврале 2019 [14]. Именно она была первой выложена в открытый доступ, в связи с опасениями компании в недобросовестном использовании моделей. Однако затем OpenAI выложила и другие модификации GPT-2. [15][16][17][18][19]

4 Методы улучшения качества генерируемого текста

В работе использовались несколько методов улучшения качества текста, среди них: использование переводчика для англоязычной модели, дообучение модели и замена слов на синонимы из тематической модели. Более подробно эти методы были рассмотрены во время проведения общей исследовательской работы. [20][21][22][23][24]

5 Данные

Сбор данных для улучшения моделей состоял из двух частей: сбор банковского корпуса и сбор тематического словаря для описанной ранее модели замен. В качестве тематики была выбрана банковская по нескольким причинам:

1) наличие специфической лексики, которая позволяла наглядно продемонстрировать качество генерируемого текста;
2) наличие большого количества текстовых ресурсов и материалов по данной тематике в сети Интернет, что позволило ускорить процесс сбора данных для моделей;
3) рекомендация руководителя НИР.

5.1 Подготовка банковского корпуса

Данный корпус использовался для дообучения нескольких моделей и составлялся по следующей схеме:

1) Выбор тематического ресурса, содержащего большое количество текстовой информации, написанной квалифицированными людьми в однообразном виде. Выбор пал на известный портал Банки.ру[1].
2) Группировка текста, полученного с данного ресурса в единый корпус.
3) Приведение текста в формат, удобный для моделей, а именно очистка текста от ссылок, эмодзи и прочих незначимых элементов языка.

Код, выполняющий процедуру сбора корпуса также опубликован и может быть найден[2] в среде Google Colab.

5.2 Подготовка тематического словаря

[1] https://www.banki.ru/
[2] https://colab.research.google.com/drive/1W3CNRWaOXCjv9l8XWqyVwOiW8wpgUz-u

Для начала корпус, описанный в п.5.1, был отфильтрован: теперь представлял из себя всевозможные встречавшиеся в оригинальном корпусе слова, приведенные в начальную форму, без предлогов, знаков препинания, междометий и других частей речи, не несущих смысловую нагрузку, а выполняющих лишь связующую функцию в предложениях. В фильтрованный корпус попадали только существительные и прилагательные.

Для составления банковского словаря, за основу был взят готовый тематический словарь [25], который вручную дополнялся различными формами имеющихся в нем слов из фильтрованного корпуса. Так, например, для слова “эквайринг” в финальный словарь было добавлено слово “эквайринговый”.

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.

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[3].

Вторая часть эксперимента заключается определении статистических гипотез, оценке сгенерированных текстов по нескольким критериям и статистической проверке гипотез при помощи различных критериев. Оценивание качества текстов более подробно было рассмотрено при проведении общей исследовательской работы.

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

В таблице 2 приведены средние значения полученных оценок.

Таблица 2 – Средние оценки текстов, сгенерированных в процессе эксперимента[4]

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

Для более формальной оценки результатов исследования было проведено статистическое исследование со следующей системой гипотез, далее мы будем использовать следующие обозначения:

H_{0n} - n-ая основная гипотеза.
H_{an} - n-ая альтернативная гипотеза

Введем следующие гипотезы:

[4] 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

H_{01} - грамматическая корректность дообученной на банковском корпусе модели с алгоритмом замен слов на схожие слова банковской тематики улучшилась в сравнении с грамматической корректностью модели ruGPT-3 модели S.

H_{a1} - грамматическая корректность дообученной на банковском корпусе модели с алгоритмом замен слов на схожие слова банковской тематики не улучшилась в сравнении с грамматической корректностью модели ruGPT-3 модели S.

H_{02} - соответствие темы дообученной на банковском корпусе модели с алгоритмом замен слов на схожие слова банковской тематики улучшилось в сравнении с критерием соответствия темы модели GPT-2 модели S.

H_{a2} - соответствие темы дообученной на банковском корпусе модели с алгоритмом замен слов на схожие слова банковской тематики не улучшилось в сравнении с критерием соответствия темы модели GPT-2 модели S.

H_{03} - соответствие темы дообученной на банковском корпусе модели с алгоритмом замен слов на схожие слова банковской тематики улучшилось в сравнении с критерием соответствия темы модели GPT-3 модели S.

H_{a3} - соответствие темы дообученной на банковском корпусе модели с алгоритмом замен слов на схожие слова банковской тематики не улучшилось в сравнении с критерием соответствия темы модели GPT-3 модели S.

H_{04} - логическая связность дообученной на банковском корпусе модели с алгоритмом замен слов на схожие слова банковской тематики улучшилась в сравнении со сгенерированным текстом модели GPT-2 модели S.

H_{a4} - логическая связность сгенерированного текста на дообученной на банковском корпусе модели с алгоритмом замен слов на схожие слова банковской тематики улучшилась в сравнении со сгенерированным текстом модели GPT-2 модели S.

H_{05} - логическая связность сгенерированного текста на дообученной на банковском корпусе модели с алгоритмом замен слов на схожие слова банковской тематики улучшилась в сравнении со сгенерированным текстом модели GPT-2 модели S.

H_{a5} - логическая связность сгенерированного текста на дообученной на банковском корпусе модели с алгоритмом замен слов на схожие слова банковской тематики не улучшилась в сравнении со сгенерированным текстом модели GPT-3 модели S.

Приступим к анализу полученных результатов. Все результаты находятся в приложении А. Проверив первую гипотезу критерием Уилкоксона, было получено эмпирическое значение $T_{эмп}$=120 при критическом значении $T_{крит}$=151 [27] (для уровня значимости a=0,05), что свидетельствует о том, что на данном уровне значимости сдвиг в типичную сторону преобладает и позволяет нам отвергнуть гипотезу H_{01}.

Во время расчета эмпирического значения для второй гипотезы было получено следующее значение: $T_{эмп}$ = 234. Размер выборки и уровень значимости на протяжении серии экспериментов не менялся, поэтому на принятом уровне значимости гипотеза H_{02} отклоняется в пользу альтернативной.

Проверив 3-ую гипотезу, было получено $T_{эмп}$ = 100. Что позволяет принять основную гипотезу и говорит о том, что построенная модель дала заметный прирост по критерию удержания темы генерируемого текста.

Оставшиеся две гипотезы относятся к последнему критерию: логическая связность генерируемого текста. В случае гипотезы №4 было получено эмпирическое значение $T_{эмп}$ = 67, что позволяет принять основную гипотезу. Во время проверки гипотезы №5 было эмпирическое значение $T_{эмп}$ = 117, что также позволяет принять основную гипотезу. Тогда итоговая таблица №3 выглядит следующим образом(зеленым цветом обозначены принятые гипотезы, красным отмечены отвергнутые гипотезы):

Таблица 3 – результаты экспериментального исследования

H_1	H_{01}	H_{a1}
H_2	H_{02}	H_{a2}
H_3	H_{03}	H_{a3}
H_4	H_{04}	H_{a4}
H_5	H_{05}	H_{a5}

7.2 Оценка достоверности полученных результатов

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[5].

[5] 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 ПРОВЕРКИ ГИПОТЕЗ КРИТЕРИЕМ УИЛКОКСОНА

Результаты проверки гипотез критерием уилкоксона приведены в таблицах А.1-А.5.

В первых двух столбцах таблиц данного приложения приведены оценки результатов генерации соответствующих моделей по соответствующим характеристикам.

Таблица А.1 – Результаты применения критерия Уилкоксона к моделям ruGPT-3 и дообученной GPT2 при проверке гипотезы H_0

ruGPT-3	Дообученная модель GPT2 на банковских текстах + алгоритм замены на схожие слова банковской тематики	Разность	Модуль разности	Ранги
4	5	-1	1	24
5	5	0	0	13
4	4	0	0	13
5	4	1	1	24
4	4	0	0	13
5	4	1	1	24
5	5	0	0	13
5	5	0	0	13
4	5	-1	1	24
4	4	0	0	13
4	5	-1	1	24
4	4	0	0	13
4	4	0	0	13
5	5	0	0	13
5	5	0	0	13
4	4	0	0	13
5	3	2	2	29
4	5	-1	1	24
5	3	2	2	29

5	5	0	0	13
5	4	1	1	24
5	5	0	0	13
5	4	1	1	24
4	5	-1	1	24
5	4	1	1	24
5	4	1	1	24
5	5	0	0	13
5	3	2	2	29
5	4	1	1	24
5	5	0	0	13

Таблица А.2 – Результаты применения критерия Уилкоксона к моделям GPT-2 и дообученной GPT2 при проверке гипотезы H_1

GPT-2	Дообученная модель GPT2 на банковских текстах + алгоритм замены на схожие слова банковской тематики	Разность	Модуль разности	Ранги
3	4	-1	1	17,5
0	4	-4	4	3,5
4	3	1	1	29,5
2	5	-3	3	7,5
3	3	0	0	25
2	3	-1	1	17,5
1	3	-2	2	11,5
5	4	1	1	29,5
4	4	0	0	25
2	4	-2	2	11,5
4	4	0	0	25
1	4	-3	3	7,5
3	3	0	0	25
2	3	-1	1	17,5

3	3	0	0	25
2	3	-1	1	17,5
1	4	-3	3	7,5
3	4	-1	1	17,5
2	3	-1	1	17,5
3	3	0	0	25
3	4	-1	1	17,5
2	4	-2	2	11,5
1	5	-4	4	3,5
2	2	0	0	25
0	3	-3	3	7,5
0	5	-5	5	1
1	3	-2	2	11,5
0	4	-4	4	3,5
1	5	-4	4	3,5
3	4	-1	1	17,5

Таблица А.3 – Результаты применения критерия Уилкоксона к моделям ruGPT-3 и дообученной GPT2 при проверке гипотезы H_2

ruGPT-3	Дообученная модель GPT2 на банковских текстах + алгоритм замены на схожие слова банковской тематики	Разность	Модуль разности	Ранги
3	4	-1	1	13,5
3	4	-1	1	13,5
3	3	0	0	3,5
4	5	-1	1	13,5
1	3	-2	2	25
5	3	2	2	25
3	3	0	0	3,5
4	4	0	0	3,5
2	4	-2	2	25

3	4	-1	1	13,5
3	4	-1	1	13,5
2	4	-2	2	25
4	3	1	1	13,5
3	3	0	0	3,5
4	3	1	1	13,5
2	3	-1	1	13,5
2	4	-2	2	25
3	4	-1	1	13,5
3	3	0	0	3,5
1	3	-2	2	25
4	4	0	0	3,5
3	4	-1	1	13,5
3	5	-2	2	25
1	2	-1	1	13,5
4	3	1	1	13,5
3	5	-2	2	25
4	3	1	1	13,5
3	4	-1	1	13,5
1	5	-4	4	30
2	4	-2	2	25

Таблица А.4 – Результаты применения критерия Уилкоксона к моделям GPT-2 и дообученной GPT2 при проверке гипотезы H_3

GPT-2	Дообученная модель GPT2 на банковских текстах + алгоритм замены на схожие слова банковской тематики	Разность	Модуль разности	Ранги
2	4	-2	2	25,5
1	3	-2	2	25,5
3	4	-1	1	15,5
2	4	-2	2	25,5

3	3	0	0	4,5
2	4	-2	2	25,5
2	3	-1	1	15,5
4	3	1	1	15,5
3	4	-1	1	15,5
3	3	0	0	4,5
3	4	-1	1	15,5
3	4	-1	1	15,5
5	5	0	0	4,5
4	4	0	0	4,5
4	5	-1	1	15,5
5	5	0	0	4,5
4	4	0	0	4,5
4	5	-1	1	15,5
3	4	-1	1	15,5
5	4	1	1	15,5
4	5	-1	1	15,5
2	4	-2	2	25,5
2	4	-2	2	25,5
5	5	0	0	4,5
3	4	-1	1	15,5
3	4	-1	1	15,5
3	4	-1	1	15,5
1	4	-3	3	29,5
3	3	0	0	4,5
2	5	-3	3	29,5

Таблица А.5 – Результаты применения критерия Уилкоксона к моделям ruGPT-3 и дообученной GPT2 при проверке гипотезы H_4

ruGPT-3	Дообученная модель GPT2 на банковских текстах + алгоритм замены на схожие слова	Разность	Модуль разности	Ранги

	банковской тематики			
4	4	0	0	5,5
3	3	0	0	5,5
3	4	-1	1	17,5
4	4	0	0	5,5
2	3	-1	1	17,5
5	4	1	1	17,5
5	3	2	2	27
4	3	1	1	17,5
4	4	0	0	5,5
2	3	-1	1	17,5
3	4	-1	1	17,5
2	4	-2	2	27
3	5	-2	2	27
4	4	0	0	5,5
4	5	-1	1	17,5
2	5	-3	3	30
4	4	0	0	5,5
4	5	-1	1	17,5
3	4	-1	1	17,5
3	4	-1	1	17,5
5	5	0	0	5,5
4	4	0	0	5,5
3	4	-1	1	17,5
5	5	0	0	5,5
4	4	0	0	5,5
3	4	-1	1	17,5
3	4	-1	1	17,5
2	4	-2	2	27
2	3	-1	1	17,5
3	5	-2	2	27