NLP в государственных службах: разработка модели генерации текста для эффективного взаимодействия с гражданами

Писаренко Антон, Романов Сергей

Май 2024

Аннотация

В рамках данной работы представлена задача разработки и обучения модели генерации текста, которая автоматизирует процесс создания ответов от представителей государственных органов на обращения граждан. Для решения этой задачи используется модель на основе архитектуры GPT-2 (Generative Pre-trained Transformer 2), которая дополнительно дообучается на специализированном наборе данных, включающем пары вопросов и ответов, собранных из реальной переписки между гражданами и представителями власти.

Ключевыми аспектами проекта являются подготовка и предобработка данных, обеспечивающие качественное обучение модели, включая кодирование типов сообщений и адаптацию модели для обработки конкретных запросов. Оценка качества генерируемых ответов осуществляется с помощью метрик BLEU, ROUGE и METEOR. В качетсве подходов генерации текста рассмотрены подходы: argmax, temperature Sampling, top-k sampling и top-p (nucleus) sampling, каждый из которых предоставляет различные уровни разнообразия и предсказуемости текста.

Ссылка на репозиторий: `https://github.com/AntonSHBK/NLP_course`.

1 Введение

Развитие ИИ оказало заметное влияние на многие аспекты жизни человека, предоставив возможности для улучшения качества и доступности услуг, ускорения и оптимизации процессов принятия решений и внедрения автоматизированных систем в различные сферы деятельности. Одной из ключевых областей, демонстрирующих потенциал ИИ, является NLP, технология, позволяющая машинам понимать, интерпретировать и генерировать человеческий язык в его естественной форме.

Искусственный интеллект преобразует промышленность и социальные процессы, делая возможным автоматизацию задач, которые ранее требовали человеческого вмешательства. Это включает в себя такие области,

как здравоохранение, где ИИ используется для диагностики заболеваний с высокой точностью, образование, где персонализированные учебные системы предлагают индивидуальные подходы к обучению, и транспорт, где автономные транспортные средства обещают сделать передвижение более безопасным и эффективным. Во всех этих случаях ИИ способствует повышению производительности, уменьшению ошибок и оптимизации ресурсов, что в конечном итоге ведет к более высокому качеству жизни и устойчивому развитию общества.

1.1 Авторы

Работа выполнена группой разработчиков:

Романов Сергей сформировал и подготовил датасет, определили целевые метрики, участвовал в модификации и отладке процееса обучения модели.

Писаренко Антон подготовил этот документ, работал над модификацией модели, реализовал визуализацию обучения, анализировал результат.

1.2 Определение области и цели проекта

В рамках данной работы представлена задача разработки и обучения модели генерации текста, которая автоматизирует процесс создания ответов от представителей государственных органов на обращения граждан. Центральной целью проекта является разработка системы, способной анализировать текстовые сообщения, поступающие от пользователей, и генерировать адекватные, информативные ответы, соответствующие заданным критериям качества и релевантности.

Для достижения этой цели необходимо выполнить следующие задачи:

1. **Определить архитектуру и конткретную базовую модель трансформера**, выбрав наиболее подходящую из доступных предобученных моделей, в зависимости от их способности к адаптации под специфические требования задачи.

2. **Модифицировать архитектуру для учета заданных параметров**, включая интеграцию дополнительных данных, таких как категория обращения или предыдущие взаимодействия пользователя с государственными службами, для повышения точности и персонализации ответов.

3. **Формирование и описание набора данных**, который будет использоваться для обучения модели. Необходимо собрать, очистить и структурировать данные, состоящие из вопросов и ответов между гражданами и государственными учреждениями, учитывая различные аспекты, такие как тип сообщения, региональные особенности и предмет обращения. Данные должны быть размечены для обучения с учетом контекста и специфики задачи.

4. **Установить и определить метрики для оценки адекватности генерирования текста**, а также разработать методику их применения для оценки как точности, так и естественности текстов.

5. **Установить способы генерации текста**, определив и интегрировав различные стратегии для генерации более качественных и разнообразных текстовых ответов.

6. **Обучить модель**, проведя тренировку на собранных и обработанных данных с целью достижения оптимальной производительности и точности ответов.

7. **Сравнить полученные результаты с эталонными ответами** (предыдущими моделями) для оценки прогресса и эффективности новой системы, используя установленные метрики.

2 Обзор литературы

Обработка естественного языка (NLP) в последние годы достигла значительных успехов благодаря развитию глубоких нейронных сетей и массовому накоплению текстовых данных. Современные NLP-системы способны не только анализировать тексты с точки зрения грамматики и синтаксиса, но и извлекать смысловые и эмоциональные составляющие, что делает их приложения чрезвычайно широкими [Pais et al., 2022]. Примеры включают автоматическую генерацию текстов, машинный перевод, создание чат-ботов для обслуживания клиентов и многое другое. Такие технологии, как трансформеры и предобученные модели вроде GPT [Radford et al., 2018] и BERT [Devlin et al., 2019], значительно продвинули понимание и генерацию естественного языка, демонстрируя впечатляющие результаты в таких задачах, как ответы на вопросы, автоматическое резюмирование и персонализированная коммуникация.

В направлении развития систем взаимодействия органов исполнительной власти и граждан, ИИ позволяет автоматизировать рутинные процедуры, ускоряя обработку запросов граждан, повышая точность административных решений и улучшая доступ к публичной информации [Ferreira, 2023]. Это способствует не только оптимизации работы государственных структур, но и усилению прозрачности и открытости власти.

Обработка естественного языка играет ключевую роль в автоматизации взаимодействий между гражданами и государственными службами. NLP-технологии позволяют разрабатывать системы, способные анализировать обращения граждан, автоматически генерировать ответы на стандартные вопросы и даже проводить первичный анализ сложных запросов, требующих вмешательства специалистов. Примеры применения включают:

- **Чат-боты и виртуальные помощник**и: Автоматизация первичной поддержки граждан, предоставление ответов на часто задаваемые вопросы, помощь в заполнении форм и подаче заявлений.

- **Анализ обратной связи**: Использование NLP для анализа писем, жалоб и предложений граждан, что помогает выявлять общие тренды и проблемные области, требующие внимания.

- **Автоматизация документооборота**: Преобразование неструктурированного текста в структурированную форму, автоматическая категоризация документов, что снижает ручной труд и повышает эффективность процессов.

Интеграция ИИ и NLP в государственные структуры несет не только возможности, но и вызовы, особенно в области этики и защиты данных. Важно обеспечивать защиту личной информации, предотвращать предвзятость в алгоритмах и разрабатывать системы таким образом, чтобы они были понятны и прозрачны для граждан [Henderson et al., 2017, Corbett-Davies et al., 2023].

На текущий момент исследования в области генерации текстов для специфических задач продемонстрировали значительный прогресс благодаря развитию технологий машинного обучения, особенно глубокого обучения. Модели, основанные на архитектурах трансформеров, таких как GPT (Generative Pre-trained Transformer) [Radford et al., 2018] и BERT (Bidirectional Encoder Representations from Transformers) [Devlin et al., 2019], выдвинулись на передний план в этой области.

Исследования, направленные на генерацию текстов различной сложности и направленности, включают создание автоматических систем отчетности, генерацию новостей, автоматизацию написания кода, составление медицинских отчётов и многое другое. Такие системы требуют не только точного воспроизведения языковых структур, но и способности адаптироваться к специфическим требованиям домена.

Одной из выдающихся работ в этом направлении является статья Васвани и других соавторов - "Attention is All You Need"[Vaswani et al., 2023], которая представила модель Трансформер (Transformer) рис. 1, лежащую в основе многих последующих исследований в области NLP. Эта архитектура позволила улучшить качество генерации текста за счёт лучшего улавливания контекста на длинных дистанциях.

Примером применения специфических генеративных моделей является исследование в области медицины, где ИИ используется для создания клинических записей и отчётов на основе данных пациентов. В работе Гуанксионга и соавторов [Liu et al., 2019] исследователи разработали модель, которая автоматически генерирует описания рентгеновских снимков грудной клетки, демонстрируя высокую клиническую точность.

В сфере программирования набирают популярность инструменты, такие как GitHub Copilot, основанные на модели GPT-3 от OpenAI [GitHub, 2021], которые могут автоматически генерировать код по запросу пользователя, облегчая процесс разработки программного обеспечения.

Ещё одним важным направлением исследований является улучшение этических аспектов генерации текста. Работа Хендерсона, Шина и дру-

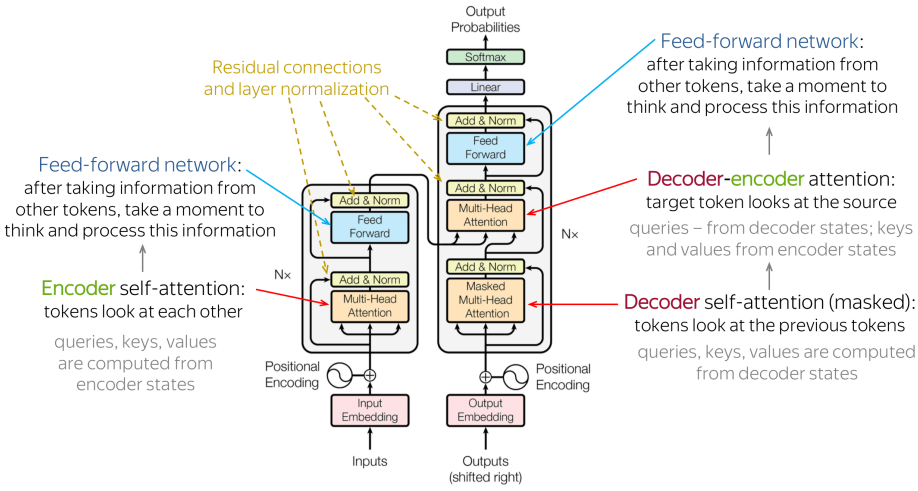

Рис. 1: Базовая модель трансформера из оригинальной статьти "Attention is All You Need"[Vaswani et al., 2023]

гих [Henderson et al., 2017] затрагивает вопросы предвзятости и транспарентности в автоматизированных системах генерации текста, подчеркивая необходимость разработки решений, способных обеспечить справедливое и беспристрастное использование ИИ в общественных и частных секторах.

Интеграция текстовых данных с дополнительными метаданными становится всё более популярной в современных исследованиях в области обработки естественного языка (NLP). Это позволяет моделям более эффективно понимать контекст и улучшать качество генерации или классификации текста. Примером такой работы является исследование Кескара [Keskar et al., 2019], в которой которой рассматривается генерация текста с учётом контролируемых атрибутов, таких как настроение и стилистика, используя модель GPT-2 для создания текста с заданными характеристиками. Также стоит отметить ещё одну работу группы авторов [Sanh et al., 2018], в которой рассматривается многоуровневый подход к обучению представлений, обучая модель одновременно на нескольких семантических задачах, что позволяет лучше улавливать семантические связи между различными типами данных.

Исследование By и Xe [Wu and He, 2019]. показывает, как включение информации о категории сущностей в архитектуру BERT улучшает понимание контекста и точность модели в задачах классификации отношений. В ещё одной работе [Martins et al., 2019] автор рассматривает совместное обучение распознавания именованных сущностей и связывания этих сущностей с их идентификаторами в базе данных, что также демонстрирует значительное улучшение процесса распознавания и связывания.

Различные научные работы и доклады подробно исследуют, как государственные органы могут использовать NLP для улучшения коммуникаций и

автоматизации процессов, что позволяет повысить качество взаимодействия с гражданами и обработку больших объемов данных.

В работе Джао и других авторов [Reis et al., 2019] рассматриваются вопросы применения ИИ в управлении и интеграции в государственные органы, а также использование NLP для автоматизации ответов на запросы граждан и улучшения доступности информации.

В другой работе группы авторов во главе Гиовани [Liva et al., 2020] рассматривается важность цифровой трансформации в государственном управлении. Авторы делают акцент на использовании NLP для оптимизации обработки запросов и предоставления более качественных услуг.

Исследование Покхреда и других авторов [Pokhrel et al., 2019] освещает потенциал ИИ и NLP в улучшении интерактивности между гражданами и государством. Основное внимание здесь уделено упрощению доступа к информации и госуслугам через улучшенные технологии обработки языка.

В России исследования в области обработки естественного языка (NLP) активно развиваются благодаря усилиям как академических, так и коммерческих организаций. Научные исследователи и разработчики применяют современные методы машинного обучения и глубокого обучения для решения широкого спектра задач, связанных с автоматизацией обработки текстов на русском языке [Mitkov and Angelova, 2021].

Одним из заметных достижений является разработка русскоязычной версии бенчмарка SuperGLUE группой авторов [Fenogenova et al., 2022]. Этот документ описывает адаптацию знаменитого англоязычного теста для оценки моделей NLP, что способствует улучшению качества и эффективности русскоязычных NLP-систем.

В Высшей Школе Экономики осуществляются проекты по анализу эмоциональной окраски текстов, что важно для мониторинга социальных медиа и анализа потребительских отзывов. Московский государственный университет фокусируется на создании и анализе больших текстовых корпусов, что помогает улучшить технологии обработки русского языка и его применение в различных областях. НИУ ВШЭ разрабатывает системы для автоматического извлечения и анализа информации из новостных потоков, способствующие выявлению общественных трендов и изменений в общественном мнении. Сбербанк активно внедряет чат-боты на основе NLP для улучшения взаимодействия с клиентами, что способствует повышению качества обслуживания и оптимизации процессов.

Эти и многие другие исследования показывают динамичное развитие области генерации текстов и важность интеграции технических, этических и практических аспектов для создания надежных и функциональных систем на базе искусственного интеллекта.

3 Сбор и подготовка данных

Данный датасет собран из открытых источников, в частности, с официального аккаунта Администрации Губернатора и Правительства Московской об-

ласти на платформе ВКонтакте [Administration of Government of MR, 2023]. Основу датасета составляют тексты сообщений пользователей, размещённые на постах и в комментариях к публикациям администрации, а также ответы от представителей государственных органов или соответствующих компетентных структур. Сбор данных осуществлялся с использованием специально разработанного краулера, который автоматизировал процесс экстракции текстовых данных.

Исходная информация (полученная краулером) имеет только: автора сообщения гражданина, текст сообщения, автора заинтересованной структуры, текст ответа. Другие параметры данных получают путём предварительной обработки данных, куда входит: определение тематики сообщения, тип сообщения, определение ответственной организации, категорию сообщения, определение региона.

Все эти параметры помечаются (разметка данных) в автоматизированном режиме, за исключением ответственного. Данные о принадлежности ответственных персон к определённым структурным подразделениям стало возможным благодаря предоставленной информации коллеги из администрации Московской области. Адрес назначается по контекстным данным, указанным в сообщении. Категория, тип и тематика сообщения размечаются обученными языковыми моделями (BERT-подобными), точность моделей варьируется от 70 - 85%.

Датасет содержит информацию о взаимодействиях граждан с представителями госорганов, 8 параметров. Распределение по категориям и типа показано на рисунках 2 и 3. Детальное описание структуры данных:

1. **responsible_person** - Персона или организация, ответственная за решение проблемы (например, "Администрация Химки").

2. **type_problem** - Тип проблемы, которую необходимо решить (например, "Устранение проблемы").

3. **topic** - Тема обращения, описывает конкретную проблему или вопрос (например, "Неудовлетворительное качество товара, оказания услуг").

4. **categoria** - Категория вопроса или проблемы (например, "Торговля, товары и услуги").

5. **region** - Регион, откуда поступило обращение (например, "Орехово-Зуевский").

6. **source** - Источник обращения или контекст, в котором возник вопрос. Содержит описание ситуации, которая привела к обращению (например, текст о цифровизации услуг).

7. **target** - Ответ уполномоченного лица госорганов на обращение гражданина. Содержит текстовый ответ на поставленный вопрос или проблему.

8. **context** - Дополнительный контекст обращения,содержит всю переписку участников диалога.

Датасет содержит только текстовую информацию (текст сообщений, метки, категории), включает 11475 записей, из которых 11140 содержат полную информацию без пропусков. Данные записаны в формате csv, размер файла 6452 КБ. К данным не применялась предварительная обработка данных (очистка). Данные разделены на тренировочный и валидационный наборы, последний составляет 10% от общего объёма данных, при этом данные были случайно перемешаны.

В сообщениях присутствует дополнительная информация, такая как ссылки, смайлики и упоминания пользователей (id пользователей которых цитировали или упоминали), которая может создавать шум. Помимо этого в данных присутствует дублирование сообщений пользователей, это связано с тем, что на одно сообщение необходим ответ нескольких ответственных лиц, имеющих различную юрисдикцию в решении проблема автора.

Анализируя распределение по категориям и типа представленны на рисунках 2 и 3 можно сделать вывод о неравномерном распределении данных по категориям, такая тенденция также наблюдается и при распределении по регионам и по ответсвенным.

Эти факторы следует учитывать при обработке и анализе данных. Представленные данные использованы в обучении и валидации модели.

Таблица 1: Общее представление данных

	responsible person	**type problem**	**topic**	**categoria**
Количество	11475	11475	11475	11475
Уникальные	181	11	120	11
Наиболее частые	Алексеев Алексей	Устранение проблемы	-	ЖКХ
Частота	717	8365	1663	4848
	region	**source**	**target**	**context**
Количество	11475	11150	11465	11475
Уникальные	56	5534	9821	5694
Наиболее частые	Другие регионы	Это ситуация в доме 15/2 - результат...	Здравствуйте! Спасибо за Ваш вопрос...	[id4847589| Александр], кто ответит за нанесён...
Частота	2513	24	111	24

4 Выбор технологии и архитектуры модели

4.1 Описание модели и её модификации

В качестве базовой модели было принято решение использовать архитектуру GPT-2 [Radford et al., 2019], предварительно облученную модель от автора ai-forever [Zmitrovich et al., 2023], предназначенная для работы с русским языком. Этот выбор обусловлен высокой адаптируемостью модели к задачам

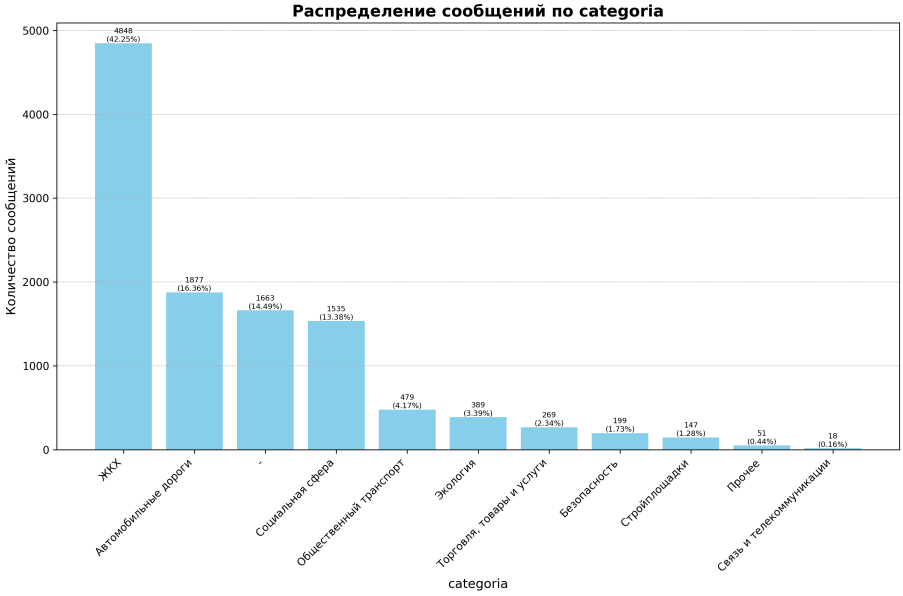

Рис. 2: Распределение по категориям

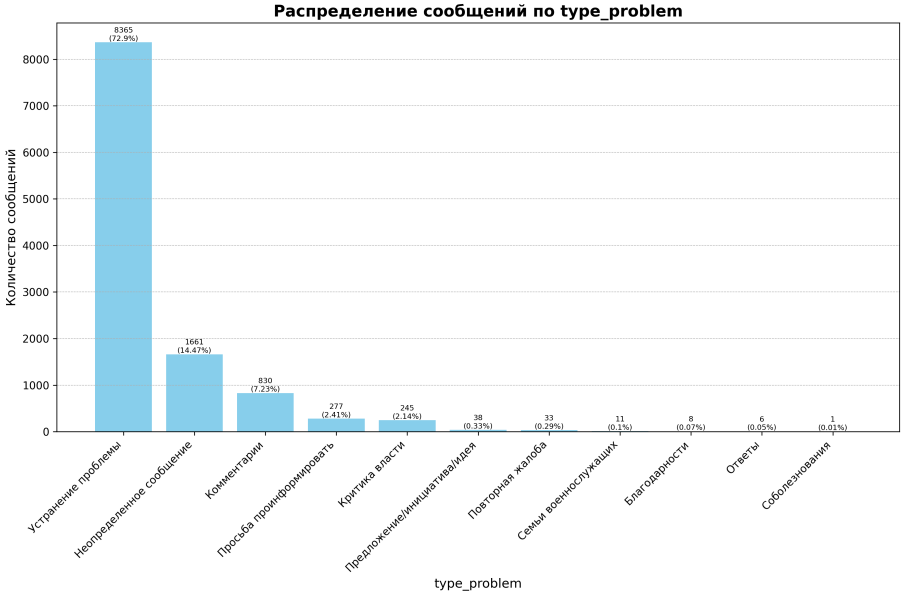

Рис. 3: Распределение по типам сообщений

генерации текста и её способностью обрабатывать большие объёмы информации для создания качественных текстовых ответов. Предварительное обучение на обширном текстовом корпусе обеспечивает модели прочную базу для дальнейшей настройки и дообучения под специфические задачи.

Разрабатываемая модель представляет собой модифицированную версию стандартной архитектуры GPT-2, адаптированную для специфических нужд взаимодействия с пользователями в контексте диалога с представителями власти. Основные модификации касаются интеграции дополнительных данных о типе сообщения, что позволяет модели более точно адаптироваться к контексту запросов и предоставлять релевантные ответы.

В качестве фреймворка проектирования модели используется PyTorch [Paszke et al., 2019].

Модель включает слой ембеддинга (nn.Embedding), что позволяет встраивать информацию о типе сообщения непосредственно в процесс обработки данных, это усиливает контекстуальное понимание модели. Дополнительный линейный слой (nn.Linear) интегрирует эмбеддинги типа сообщения с токенными эмбеддингами, обеспечивая корректный проход базовой модели. Выходной линейный слой преобразует последние скрытые состояния в логиты, необходимые для генерации последующих токенов.

В процессе настройки модели было принято решение о заморозке весов основной части модели GPT-2, за исключением последних двух голов. При проведении серии экспериментов было выявлено что такой подход благоприятно сказываетс на результате.

В функции передачи данных ‘forward‘, проектируемой модели, процесс начинается с генерации эмбеддингов для каждого токена и типа сообщения. Эти эмбеддинги затем комбинируются и усиливаются дополнительным линейным слоем (размер слоя равен размеру скрытого слоя базовой модели. 768), известным как. После этого, комбинированные эмбеддинги подаются в основную часть модели GPT-2. Модель обрабатывает входные данные, учитывая маску внимания, что позволяет модели сосредоточиться на релевантных частях входной последовательности. В завершающем этапе, последние скрытые состояния, полученные от GPT-2, преобразуются в логиты с помощью выходного линейного слоя (размер слоя равен размеру словаря). Эти логиты представляют собой вероятности следующих токенов, которые модель использует для генерации текста, обеспечивая по идее тем самым точность и релевантность генерируемых ответов.

В качестве оптимизатора модели для обучения был выбран оптимизатор AdamW [Loshchilov and Hutter, 2018], который является модификацией традиционного алгоритма Adam [Kingma and Ba, 2014]. AdamW вносит улучшения в обработку штрафов за регуляризацию, что помогает лучше контролировать веса в сети и предотвращает их чрезмерный рост, обеспечивая более стабильное и эффективное обучение.

В качестве функции потерь была выбрана CrossEntropyLoss [Goodfellow et al., 2016], которая широко используется для задач классификации с множественными классами. Эта функция потерь оценивает, насколько вероятности, предсказанные моделью для каждого класса, соответствуют фактическим меткам

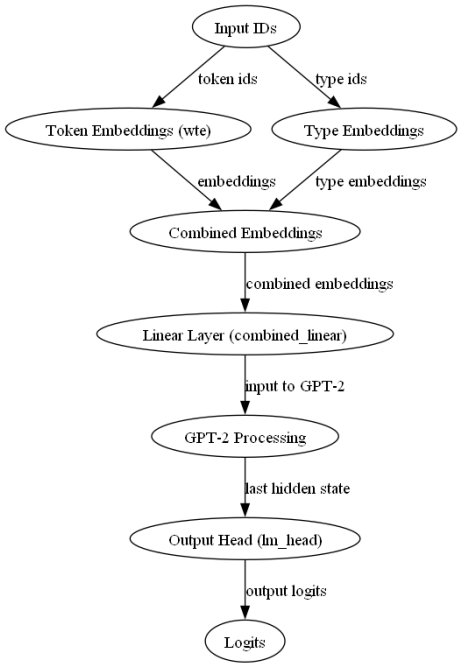

Рис. 4: Общая структура моедли

класса.

4.2 Метрики валидации

Для оценки качества генерируемых текстов рассмотрим использование трёх широко распространённых метрики: BLEU [Papineni et al., 2002], ROUGE [Lin, 2004] и METEOR [Banerjee and Lavie, 2005]. Эти метрики позволяют количественно анализировать соответствие сгенерированных ответов эталонным ответам и оценивать их по различным аспектам качества, таким как точность, покрытие и упорядоченность.

4.2.1 BLEU (Bilingual Evaluation Understudy)

BLEU — это одна из наиболее популярных метрик для оценки качества машинного перевода, которая также широко применяется для задач генерации текста. BLEU измеряет, насколько n-граммы сгенерированного текста совпадают с n-граммами в эталонных текстах, учитывая их частоту вплоть до заданного размера n. BLEU оценивает точность, но с поправкой на «штраф за длину», чтобы избежать чрезмерно кратких ответов, которые искусственно могли бы увеличить совпадение n-грамм [Papineni et al., 2002].

BLEU оценка рассчитывается следующим образом:

1. **Совпадение n-грамм**: Для каждой n-граммы в сгенерированном тексте проверяется, встречается ли она в эталонном тексте. Для каждой n-граммы вычисляется отношение числа совпадений к общему числу n-грамм в сгенерированном тексте.

$$Precision_n = \frac{\sum_{\text{n-gram} \in \text{Candidate}} \min(\text{Count(n-gram)}, \text{MaxRefCount(n-gram)})}{\sum_{\text{n-gram} \in \text{Candidate}} \text{Count(n-gram)}}$$

где $MaxRefCount(n-gram)$ — максимальное количество данной n-граммы среди всех эталонных текстов, $Count(n-gram)$ — количество данной n-граммы в кандидате.

2. **Геометрическое среднее**: Вычисляется геометрическое среднее из точностей n-грамм для различных n.

3. **Штраф за короткие тексты (Brevity Penalty, BP)**: Если сгенерированный текст короче эталонного, вводится штраф за короткую длину для предотвращения предпочтения необоснованно коротких ответов.

$$BP = \begin{cases} 1 & \text{если } c > r \\ e^{(1-r/c)} & \text{если } c \leq r \end{cases}$$

где c — длина сгенерированного текста, а r — длина эталонного текста или средняя длина нескольких эталонных текстов.

Итоговая оценка BLEU вычисляется как произведение геометрического среднего точности по всем n-граммам на штраф за короткие тексты (BP).

4.2.2 ROUGE (Recall-Oriented Understudy for Gisting Evaluation)

ROUGE используется для оценки автоматических рефератов или переводов и сосредоточена на полноте ответа, т.е., сколько n-грамм эталонного ответа захватывает сгенерированный ответ. ROUGE-L и ROUGE-N (где N указывает на размер n-грамм) — наиболее распространённые вариации [Lin, 2004].

Формула ROUGE-N:

$$\text{ROUGE-N} = \frac{\sum_{s \in \text{Reference Summaries}} \sum_{n \in s} \text{Count}_{\text{match}}(n)}{\sum_{s \in \text{Reference Summaries}} \sum_{n \in s} \text{Count}(n)}$$

ROUGE-L фокусируется на длине наиболее длинной общей подпоследовательности, что позволяет оценить не только наличие ключевых слов и фраз, но и их последовательность в тексте, что важно для оценки качества и естественности текста.

4.2.3 METEOR (Metric for Evaluation of Translation with Explicit ORdering)

METEOR — это метрика для оценки машинного перевода, которая была разработана как альтернатива BLEU для лучшего учета качества перевода с точки зрения человеческой оценки. Она учитывает не только точное совпадение слов, но и синонимы, стемминг и порядок слов, позволяя получить более гибкую и всестороннюю оценку. В отличие от BLEU, METEOR учитывает как точность, так и полноту, вводя понятия precision (P) и recall (R), и использует их для вычисления F-меры. Кроме того, METEOR вводит понятие "штраф за непоследовательность"(penalty), учитывающее различия в порядке слов между сгенерированным текстом и эталонным [Banerjee and Lavie, 2005].

Основная формула METEOR включает в себя вычисление F-меры и штрафа за непоследовательность:

$$\text{METEOR} = (1 - \text{penalty}) \cdot F_{\text{mean}}$$

где

$$F_{\text{mean}} = \frac{10 \cdot P \cdot R}{R + 9 \cdot P}$$

где P — точность (precision), доля совпадающих слов в переводе относительно общего числа слов в сгенерированном тексте; R — полнота (recall), доля совпадающих слов в переводе относительно общего числа слов в эталонном тексте; penalty — штраф за непоследовательность, вычисляемый на основе числа и длины совпадающих фрагментов слов в сгенерированном и эталонном текстах.

Существуют и другие метрики оценки, однако остановимся на наиболее распространённых методах.

4.3 Генерация текста

В данной работе рассмотрены четыре основных метода генерации текста, которые широко используются в современных моделях генерации естественного языка. Каждый из этих методов имеет свои особенности и применяется для достижения различных аспектов разнообразия и точности в сгенерированных текстах [Holtzman et al., 2019].

4.3.1 Temperature Sampling

Temperature sampling модифицирует распределение вероятностей, делая его более "мягким"или "жестким в зависимости от значения параметра температуры T. При $T > 1$ распределение становится более равномерным, что увеличивает разнообразие генерируемых ответов. При $T < 1$ распределение становится более "острым уменьшая разнообразие и увеличивая детерминированность выбора.

Формула temperature sampling:

$$P(i) = \frac{\exp(\log(p_i)/T)}{\sum_j \exp(\log(p_j)/T)}$$

где p_i — исходная вероятность токена i.

4.3.2 Top-k Sampling

Top-k sampling ограничивает выборку следующего токена только k наиболее вероятными токенами. Этот метод уменьшает риск выбора маловероятных токенов и позволяет сосредоточиться на более вероятных вариантах, что улучшает когерентность текста при сохранении элемента случайности.

Формула Top-k Sampling: Выбирается подмножество токенов C из всех возможных токенов V, где $|C| = k$ и каждый токен из CS имеет максимальные вероятности из V. Затем выполняется:

$$P(i) = \begin{cases} \frac{p_i}{\sum_{j \in C} p_j} & \text{if } i \in C \\ 0 & \text{otherwise} \end{cases}$$

4.3.3 Top-p Sampling (Nucleus Sampling)

Top-p sampling, также известный как nucleus sampling, выбирает минимальный набор токенов C, сумма вероятностей которых составляет p. Это позволяет исключить наименее вероятные токены и сосредоточить выборку на более вероятном "ядре"распределения.

Формула Top-p Sampling: Выбираются токены так, что:

$$\sum_{i \in C} p_i \geq p$$

и выполняется нормализация вероятностей для токенов в C:

$$P(i) = \begin{cases} \frac{p_i}{\sum_{j \in C} p_j} & \text{if } i \in C \\ 0 & \text{otherwise} \end{cases}$$

5 Обучение, оценка и оптимизация модели

Дообучение модели настраивается с учётом следующих гиперпараметров:

- **Максимальная длина последовательности**: ‘max_length=64‘ обеспечивает баланс между детализацией ответов и вычислительной эффективностью;
- **Размер батча**: ‘batch_size=64‘ обеспечивает балансированную загрузку данных в модель во время обучения;
- **Размер тестового набора**: ‘test_size=0.1‘ позволяет выделить 10

- **Скорость обучения**: ‘learning_rate=1e-5‘;
- **Количество эпох**: ‘num_epochs=10‘;
- **Температура генерации**: ‘temperature=0.7‘;
- **Коэффициент top_k**: ‘top_k=11‘;
- **Коэффициент top_p**: ‘top_p=0.9‘;

Для обучения модели использовались следующие аппаратные средства:

- **Процессор**: 12th Gen Intel(R) Core(TM) i5-1240P с тактовой частотой 1.70 GHz, что обеспечивает достаточную мощность для обработки данных и расчётов;
- **Оперативная память**: 64,0 ГБ, что позволяет эффективно работать с большими объёмами данных и управлять несколькими процессами одновременно без существенной потери производительности;
- **Видеокарта**: NVIDIA RTX A6000, одна из передовых графических карт, которая поддерживает ускорение вычислений с помощью CUDA.

Обучение модели производилось с использованием библиотеки ‘PyTorch‘ и технологии ‘CUDA‘, что позволило полностью использовать вычислительные мощности графического процессора для обработки операций при обучении. Это значительно сократило время обучения и повысило его эффективность.Время обучения модели составило 21 минут и 58 секунды. Это относительно короткий период для моделей такого типа, благодаря чему проект может быть быстро адаптирован и масштабирован.После завершения обучения веса модели были сохранены для последующего использования. Это позволяет легко воспроизвести результаты и использовать обученную модель для генерации ответов без необходимости повторного обучения.

6 Анализ результатов

По итогам обучения модели, основанной на архитектуре GPT-2, можно сделать несколько ключевых наблюдений. График обучения показывает, что потери как на этапе обучения, так и валидации постепенно уменьшаются, что свидетельствует о стабилизации процесса обучения. Однако, несмотря на уменьшение потерь, анализ сгенерированных текстов выявляет значительные проблемы с качеством и релевантностью ответов.

В таблице 2 представлены примеры входных данных (тексты обращений граждан) и соответствующие им целевые результаты (ответы органов власти). Это позволяет наглядно оценить исходные задачи, которые ставились перед моделью. Таблица 3 демонстрирует результаты, сгенерированные моделью по различным методам генерации текста, включая temperature sampling, Top-k и Top-p sampling. Анализ этих результатов показывает, что

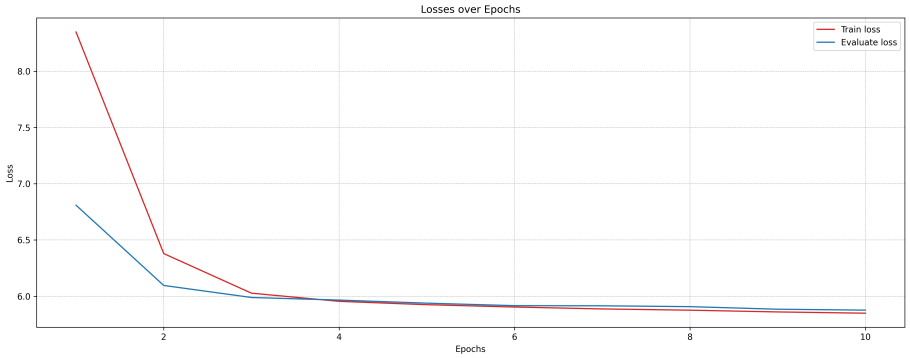

Рис. 5: График обучения, потери при обучении и валидации

сгенерированные тексты сильно отличаются от целевых. Несмотря на наличие отдельных слов из целевых текстов в сгенерированных ответах, общее качество и структура ответов оставляют желать лучшего, что указывает на недостаточную адаптацию модели к специфике задачи.

Таблица 4 содержит численные значения метрик (BLEU, ROUGE, METEOR) по проверочным данным. Низкие показатели по этим метрикам подтверждают, что модель не достигла необходимого уровня понимания и воспроизведения контекста запросов и ответов. Это подчеркивает необходимость дальнейшего анализа и возможной корректировки архитектуры модели и гиперпараметров.

Выводы, полученные в результате анализа, указывают на то, что несмотря на некоторые успехи в обучении (например, способность модели улавливать ключевые слова из запросов), генерация текста не соответствует заданным стандартам качества и релевантности. Это означает, что текущая реализация модели требует дополнительной доработки, включая возможное изменение подходов к обучению и более глубокое изучение данных, на которых обучается модель.

Разработанный алгоритм продемонстрировал свою работоспособность и потенциал для дальнейших улучшений. Несмотря на текущие трудности с релевантностью и качеством сгенерированных текстов, алгоритм успешно справился с базовой задачей генерации текста, что является важным шагом на пути к созданию полноценной системы автоматизированного ответа на обращения граждан.

Возможности для дальнейшего развития алгоритма включают его видоизменение и адаптацию под конкретные потребности и условия использования. Основываясь на анализе выбранных метрик качества, и используя различные методы генерации текста, можно добиться значительного улучшения качества и релевантности ответов модели.

Source	Target
Добрый день . Подскажите город Кашира .В деревне Хитровка когда будет назначен староста ? Был Кузнецов В.В . При нем была построена дорога , тротуар , отремонтирована бочага и многое другое ... Было голосование , люди приходили голосовать , тратили свое время . Где эти результаты ? Почему проигнорировали голоса людей ? Во всех деревнях выбрали , а на Хитровке нет . Почему ? Какое то предвзятое отношение . Почему совет депутатов не рассмотрел результаты голосования в деревне Хитровка ? Хочу получить внятный ответ на вопрос . Голосование проходило еще в конце 2023 года ! Уже в других деревнях вручают удостоверение старостам , а у нас так и не выбран !	[id398594399|Анна Соколова] добрый день! На Ваше обращение сообщаем, что по решению Совета депутатов от 26.03.2024 №21 старостой сельского населенного пункта городского округа Кашира Московской области д. Хитровка назначен Кузнецов Вячеслав Васильевич. С уважением, Администрация г.о. Кашира

Таблица 2: Соответствие между исходным текстом и целевым ответом

7 Дальнеишее направление развития исследований

В рамках дальнейшего развития проекта по генерации текстовых ответов от представителей государственных органов планируется провести ряд исследований и улучшений. Основное внимание будет уделено пересмотру и возможной модификации архитектуры модели. Это включает в себя сравнение текущей модели GPT-2 с другими подходами и архитектурами, которые могут лучше справляться с задачей учета контекста запросов и генерации более релевантных ответов.

Одним из важных направлений будет расширение и дополнительная балансировка набора входных данных. Планируется обогатить датасет более разнообразными примерами обращений, чтобы улучшить способность модели адаптироваться к различным типам запросов и минимизировать предвзятость в ответах. Это потребует тщательного анализа текущих данных и возможно, включение новых источников информации.

Также будет проведен анализ различных методов генерации текста, для определения наиболее эффективных способов в контексте задачи. Это исследование поможет найти оптимальный баланс между разнообразием и точностью ответов.

Method	Generated Text
argmax	.....
sampling	.., уважением.! в. . уважением уважением..ого8.20, г понимание .римС. . территориальной.го, свои.носим Администрация.рина указанному. месте служб.[г работу61 ...14 г можнотыным.-!.С уважением Вы.
top_k	-. уважением. уважением. уважением.. уважением, С-!.. уважением,-. уважением уважением уважением. уважением,. в-.. ., С„ уважением. уважением в, по уважением... в уважением. -.
top_p	. номера конкре99 Вашескую 8 и,!.овой С «.4 подроб. нетИ и уточен)71 Прос уважением вогоИн495род20руд на По уважением требования уважением800веПос....48 сотрудникамиИлагода Администрация обратной ВыЕв .С89 извинения Администрация

Таблица 3: Генерированный текст для разных методов

Method	BLEU	ROUGE	METEOR
Temperature Sampling	0	0.09	0.0
Top-k	5.62e-232	0.08	0.03
Top-p	6.81e-232	0.06	0.04

Таблица 4: Оценки BLEU, ROUGE и METEOR для разных методов генерации

Важной целью является масштабирование модели для её интеграции в реальную рабочую среду. Это включает в себя разработку и тестирование API для взаимодействия с моделью, а также подготовку инфраструктуры для её эффективного и безопасного использования в продуктивной среде. Интеграция модели в рабочий сервис позволит оценить её эффективность в реальных условиях и определить дополнительные области для улучшения.

8 Заключение

В ходе реализации проекта была разработана и обучена модель, основанная на архитектуре GPT-2, для автоматизации процесса генерации ответов на обращения граждан от представителей государственных органов. Обучение модели было успешно проведено на современной аппаратной платформе, и модель демонстрирует способность генерировать текстовые ответы. Однако, результаты анализа качества сгенерированных текстов показали, что релевантность и понятность этих ответов оставляют желать лучшего. Тексты часто получаются непонятными и не полностью соответствуют заданному контексту запросов, что свидетельствует о недостаточной адаптации модели к специфике задачи.

Эти наблюдения указывают на необходимость дальнейшей работы и более глубокого анализа модели. Возможно, потребуется пересмотреть подходы к дообучению, внести корректировки в архитектуру или параметры обучения, а также использовать более обширные или детализированные обучающие данные. Также целесообразно применить дополнительные методы оценки качества и релевантности ответов для более точной настройки модели под реальные потребности пользователей.

Таким образом, хотя первоначальные результаты показывают определённый потенциал применения модели GPT-2 для автоматизации ответов государственных органов, проект требует дополнительных исследований и улучшений.

Список литературы

[Administration of Government of MR, 2023] Administration of Government of MR (2023). Official vkontakte account of the administration of the governor and government of moscow region. `https://vk.com/pressmo`. Accessed: 2023-04-12.

[Banerjee and Lavie, 2005] Banerjee, S. and Lavie, A. (2005). Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72. Association for Computational Linguistics.

[Corbett-Davies et al., 2023] Corbett-Davies, S., Gaebler, J. D., Nilforoshan, H., Shroff, R., and Goel, S. (2023). The measure and mismeasure of fairness.

[Devlin et al., 2019] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding.

[Fenogenova et al., 2022] Fenogenova, A., Tikhonova, M., Mikhailov, V., Shavrina, T., Emelyanov, A., Shevelev, D., Kukushkin, A., Malykh, V., and Artemova, E. (2022). Russian superglue 1.1: Revising the lessons not learned by russian nlp models.

[Ferreira, 2023] Ferreira, C. (2023). A short review of the main concerns in a.i. development and application within the public sector supported by nlp and tm.

[GitHub, 2021] GitHub (2021). Introducing github copilot: your ai pair programmer. `https://github.blog/2021-06-29-introducing-github-copilot-ai-pair-programmer/`. Accessed: 2023-04-12.

[Goodfellow et al., 2016] Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. MIT Press.

[Henderson et al., 2017] Henderson, P., Sinha, K., Angelard-Gontier, N., Ke, N. R., Fried, G., Lowe, R., and Pineau, J. (2017). Ethical challenges in data-driven dialogue systems.

[Holtzman et al., 2019] Holtzman, A., Buys, J., Du, L., Forbes, M., and Choi, Y. (2019). The curious case of neural text degeneration. *arXiv preprint arXiv:1904.09751*.

[Keskar et al., 2019] Keskar, N. S., McCann, B., Varshney, L. R., Xiong, C., and Socher, R. (2019). Ctrl: A conditional transformer language model for controllable generation.

[Kingma and Ba, 2014] Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

[Lin, 2004] Lin, C.-Y. (2004). Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81. Association for Computational Linguistics.

[Liu et al., 2019] Liu, G., Hsu, T.-M. H., McDermott, M., Boag, W., Weng, W.-H., Szolovits, P., and Ghassemi, M. (2019). Clinically accurate chest x-ray report generation.

[Liva et al., 2020] Liva, G., Codagnone, C., Misuraca, G., Gineikyte, V., and Barcevičius, E. (2020). Exploring digital government transformation: a literature review. pages 502–509.

[Loshchilov and Hutter, 2018] Loshchilov, I. and Hutter, F. (2018). Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

[Martins et al., 2019] Martins, P. H., Marinho, Z., and Martins, A. F. T. (2019). Joint learning of named entity recognition and entity linking.

[Mitkov and Angelova, 2021] Mitkov, R. and Angelova, G., editors (2021). *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, Held Online. INCOMA Ltd.

[Pais et al., 2022] Pais, S., Cordeiro, J., and Jamil, M. L. (2022). Nlp-based platform as a service: a brief review. *Journal of Big Data*, 9.

[Papineni et al., 2002] Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318. Association for Computational Linguistics.

[Paszke et al., 2019] Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al. (2019). Pytorch: An imperative style, high-performance deep learning library. `https://pytorch.org`. Accessed: 2023-04-12.

[Pokhrel et al., 2019] Pokhrel, S. R., Sood, K., Yu, S., and Nosouhi, M. R. (2019). Policy-based bigdata security and qos framework for sdn/iot: An analytic approach. In *IEEE INFOCOM 2019 - IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, pages 73–78.

[Radford et al., 2018] Radford, A., Narasimhan, K., Salimans, T., and Sutskever, I. (2018). Improving language understanding by generative pre-training. *OpenAI Blog*.

[Radford et al., 2019] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. (2019). Language models are unsupervised multitask learners. In *OpenAI Blog*.

[Reis et al., 2019] Reis, J., Espírito Santo, P., and Melao, N. (2019). *Artificial Intelligence in Government Services: A Systematic Literature Review*, pages 241–252.

[Sanh et al., 2018] Sanh, V., Wolf, T., and Ruder, S. (2018). A hierarchical multi-task approach for learning embeddings from semantic tasks.

[Vaswani et al., 2023] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2023). Attention is all you need.

[Wu and He, 2019] Wu, S. and He, Y. (2019). Enriching pre-trained language model with entity information for relation classification.

[Zmitrovich et al., 2023] Zmitrovich, D., Abramov, A., Kalmykov, A., Tikhonova, M., Taktasheva, E., Astafurov, D., Baushenko, M., Snegirev, A., Shavrina, T., Markov, S., Mikhailov, V., and Fenogenova, A. (2023). A family of pretrained transformer language models for russian.