МИНОБРНАУКИ РОССИИ
Федеральное государственное бюджетное образовательное учреждение
высшего образования
«САРАТОВСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н. Г. ЧЕРНЫШЕВСКОГО»
Кафедра дискретной математики и информационных технологий

ГЕНЕРАЦИЯ ТЕКСТА НА ОСНОВЕ ЛИТЕРАТУРНЫХ ПРОИЗВЕДЕНИЙ

АВТОРЕФЕРАТ БАКАЛАВРСКОЙ РАБОТЫ

студента 4 курса 421 группы
направления 09.03.01 — Информатика и вычислительная техника
факультета КНиИТ
Громова Никиты Михайловича

Научный руководитель
ассистент _____ А. А. Трунов

Заведующий кафедрой
доцент, к. ф.-м. н. _____ Л. Б. Тяпаев

Саратов 2023

ВВЕДЕНИЕ

Темой представленной выпускной квалификационной работы является генерация текста на основе литературных произведений. Она включает в себя три главы:

1. Анализ существующих подходов к анализу текстов;
2. Описание методов генерации текста на основе литературных произведений;
3. Создание и сравнение моделей машинного обучения для генерации текстов.

Актуальность темы, рассмотренной выпускной квалификационной работе, обусловлена растущей ролью генерации текста в различных сферах деятельности, таких как бизнес, медиа, образование и наука. Генерация текста на основе литературных произведений может помочь создавать новые и уникальные тексты, которые отражают стиль и смысл оригинальных авторов. Такие тексты могут быть использованы для развлечения, обучения, рекламы и других целей. Однако, для эффективной генерации текста необходимо использовать современные методы и алгоритмы, которые учитывают лексические, синтаксические, семантические и стилистические особенности текстов. Поэтому изучение и сравнение различных методов генерации текста на основе литературных произведений является актуальной и важной задачей.

Материалами данного исследования выступают статьи различных авторов по теме работы. В основном они представлены научными статьями, рассматривающими различные способы и аспекты создания нейросетей для генерации текста.

КРАТКОЕ СОДЕРЖАНИЕ РАБОТЫ

В дипломной работе рассмотрены существующие методы анализа текста и его генерации, такие как метод марковских цепей, генеративно-состязательные сети, Seq2Seq модели, рекуррентные нейронные сети, Transformer модели и гибридные методы. Проанализированы преимущества и недостатки каждого метода, а также проблемы и ограничения при генерации текстов. Разработаны и реализованы программы для генерации текстов на основе LSTM модели, модели на основе GPT2 и марковских цепей. Создан веб-интерфейс для реализованных моделей. Проведено сравнение результатов работы алгоритмов.

В первой главе проводится анализ существующих подходов к анализу текстов:

— Анализ лексики и стиля обычно включает в себя частотность употребления отдельных слов и словосочетаний, выбор синонимов и антонимов, оценку эмоциональной окраски текста, выбора лексики и грамматической конструкции, а также использование фигур речи. Для эффективного анализа лексики и стиля необходимо использование интеллектуальных алгоритмов и технологий компьютерной лингвистики. Основными задачами компьютерного анализа лексики и стиля являются автоматическая классификация текстов, извлечение ключевых слов и фраз, определение тоновой окраски текста и автоматический перевод. В современных системах анализа лексики и стиля используются различные методы машинного обучения, в том числе нейронные сети.

— Семантический анализ – один из основных методов анализа текстов, который направлен на выявление значений слов и выражений в тексте, а также их взаимосвязей. С помощью семантического анализа можно определить тематику текста, выделить ключевые понятия и проверить соответствие содержания текста его заголовку. Одним из алгоритмов семантического анализа является выделение контекстуальных смыслов с помощью анализа окружения слова. Этот метод позволяет определить значения слова в зависимости от контекста, в котором оно используется.

— Анализ синтаксиса проводится для определения структуры предложений и выражений, которые могут содержать множество сложных конструкций, включая вложенные предложения и различные типы союзов. Для проведения синтаксического анализа в современных системах об-

работки естественного языка широко применяются методы машинного обучения, которые позволяют автоматически находить и распознавать различные структуры в текстах.

— Анализ смысла и контекста позволяет определить смысловую нагрузку текста в зависимости от содержания и направелнности. При генерации это может сыграть большую роль, поменяв смысл текста. Однако, при анализе семантики для генерации существуют такие проблемы, как полисемия, или многозначность, когда одно и то же слово может иметь несколько разных значений, в зависимости от контекста, что требует дополнительной корректировки алгоритмов, и генерация логически связного текста, что подразумевает понимание взаимосвязи слов и их логичное сочетание с предыдущими и последующими.

— Обработка естественного языка включает в себя преобразование естественного языка в структурированные данные, которые могут быть обработаны компьютером. Она включает в себя уже перечисленные методы, а так же несколько других, такие как графематический и морфологический анализ. Одной из проблем при обработке естественного языка является то, что одно и то же слово может иметь несколько значений в зависимости от контекста, в котором оно используется.

В конце главы делается обзор проблем и ограничений при генерации текстов. Главной проблемой является ограниченный характер методов генерации текста, что приводит к постоянному повторению фраз, слабому построению предложений и недостаточному разнообразию и качеству результирующих текстов. Более того, существующие методы во многих случаях не учитывают окружающий контекст и семантику текста, создавая содержание, лишенное смысла и не подходящее для решения определенных вопросов.

Во второй главе описываются методы генерации текста на основе литературных произведений:

— Метод марковских цепей – это вероятностная модель, описывающая последовательность возможных событий, которая может предсказывать результаты будующих действий, основываясь только на текущем состоянии процесса, и, что самое важное, подобные предсказания не уступают прогнозам, которые можно сделать, основываясь на полной истории всего процесса. Это один из самых старых и зарекомендовавших

себя методов генерации текста. Одним из основных преимуществ метода марковских цепей является его относительная простота и быстрая работа, что позволяет генерировать тексты даже на слабых системах. Однако, у этого метода есть и недостатки, в частности, чем проще модель, тем бессмысленнее получаются тексты, но и модели более высоких порядков порой могут выдавать совершенно бессвязные и лишенные смысла результаты. Кроме того, этот подход плохо учитывает зависимости между цепочками из нескольких слов, что может привести к неестественным текстам.

— Генеративно-состязательные сети (GAN) – это класс алгоритмов машинного обучения, который используется для генерации новых данных на основе уже существующих. Основная идея генеративно-состязательных сетей заключается в наличии двух нейронных сетей: генератора и дискриминатора, которые работают в паре. Генератор создаёт новые данные, основываясь на обучающих, а дискриминатор анализирует результаты генератора проводя сравнение с реальными образцами, и определяет степень их подлинности.

— Seq2Seq модели используются для задачи машинного перевода, где две нейронные сети – энкодер и декодер – используются для преобразования одной последовательности в другую. Модель состоит из двух основных компонентов – кодера и декодера. Модели Seq2Seq обучаются на наборе данных пар вход-выход, где вход и выход – это последовательность лексем. Модель стремиться максимизировать вероятность правильной выходной последовательности с учетом входной. Для повышения производительности в моделях Seq2Seq часто используется механизм внимания, который позволяет декодеру сосредоточиться на определенных частях входной последовательности при генерации результата. Такие механизмы позволяют декодеру обращаться к различным частям входной последовательности в зависимости от того, что он генерирует в текущий момент. Например, если декодер генерирует следующее слово, тесно связанное с каким-то конкретным словом в исходной последовательности, то механизм внимания переключает внимание декодера на это конкретное слово. Однако, Seq2Seq модели могут страдать от того, что они склонны повторяться или генерировать предсказуемую после-

довательность. Для преодоления этой проблемы, можно применять такие методы, как шумовые входные данные, температурный параметр и различные техники обучения с учителем.

— Рекуррентные нейронные сети (RNN) – это вид нейронных сетей, в которых связи между элементами образуют направленную последовательность. Эти алгоритмы обычно используются для решения таких задач, как перевод, обработка естественного языка (NLP), распознавание речи и создание подписей к изображениям. RNN работает с помощью обратной связи. Она имеет память, в которой хранятся предыдущие состояния. Они комбинируются с текущим входом, чтобы получить выход и следующее состояние. Одним из главных преимуществ RNN является то, что они могут обрабатывать данные различной длины, что позволяет им генерировать длинные тексты в различных стилях. Так же стоить отметить, что RNN хорошо работают с языком и могут учитывать сложности грамматической структуры и синтаксиса. Однако, у RNN есть свои недостатки. Например, они не всегда могут учитывать длинные зависимости в данных, из-за чего на выходе может получиться бессмыслица. Кроме того, они страдают от проблемы затухания градиента, которая происходит при обучении на длинных последовательностях. Одним из способов решения этих проблем являются модификации базовых RNN, такие как LSTM и GRU модели.

— LSTM (Long-Short Term Memory) – это тип RNN, который был разработан таким образом, что проблема затухания градиента почти полностью устранена, а модель обучения осталась неизменной. В LSTM нет необходимости хранить конечное число состояний заранее, как это требуется в модели Маркова, при этом они предоставляют больший диапазон параметров, таких как скорость обучения, смещения на входе и выходе. Следовательно, нет необходимости в тонкой настройке. Основное различие между архитектурами RNN и LSTM заключается в том, что скрытый слой LSTM представляет собой управляемый блок. Он состоит из четырех слоев, которые взаимодействуют друг с другом таким образом, чтобы на выходе кроме результата было ещё и состояние ячейки. Эти два параметра затем передаются на следующий слой.

— Transformer модели – это один из самых современных и продвинутых

методов генерации текста. Трансформеры обладает способностью решать задачу генерации текста путем моделирования соотношений между словами, фразами и предложениями на больших объёмах данных. Transformer модели используют внимание (attention) для оценки важности каждого слова в предложении при генерации следующего слова. Это позволяет сети учитывать контекст предыдущих слов и их зависимости, избегать повторений и создавать более связный и логически последовательный текст. Важной особенностью Трансформеров является их способность генерировать согласованные и разнообразные фразы и предложения на основе нескольких источников, что дополнительно расширяет возможности их применения. Одним из наиболее популярных способов использования таких моделей является генерация текстов для чат-ботов, которые могут в меру осмысленно отвечать на вопросы и даже вести беседы.

— Bidirectional Encoder Representations from Transformers (BERT) представляет собой нейронную сеть, основу которой составляет композиция кодировщиков трансформера. BERT является автокодировщиком. В каждом слое кодировщика применяется двустороннее сканирование, что позволяет модели учитывать контекст с обеих сторон от рассматриваемого токена, а значит, точнее определять значения токенов. В зависимости от конечной цели используют либо машинное обучение с учителем, либо без него.

— Generative Pre-trained Transformer 2 (GPT-2) – это достаточно современная языковая модель, разработанная OpenAI в 2019 году. Это модель только для декодера, использующая архитектуру трансформера для генерации текста. GPT-2 обучен на огромном количестве текстовых данных и может генерировать высококачественный текст, практически неотличимый от написанного человеком. GPT-2 использует входной текст для создания начального контекста и дальнейшей генерации текста. Длина входной строки может варьироваться от нескольких слов до максимальной длины последовательности в 1024 лексемы. Чем длиннее начальный входной текст, тем больше предметного контекста предоставляется модели. Модель обучается, предсказывая следующее слово с учетом всех предыдущих слов в некотором тексте. Эта задача приме-

няется рекурсивно для генерации нового текста.

— Гибридные методы подразумевают под собой использование двух и более моделей для генерации текста. В основе гибридных методов лежит идея использования сильных сторон каждого из используемых методов для создания более качественных текстов. Такие методы позволяют не только генерировать более разнообразный и приближенный к реальному текст, но и решать проблемы, связанные с переобучением и недообучением моделей.

В главе рассматриваются основные принципы, преимущества и недостатки каждого метода, а также их примеры применения для генерации текста. В конце главы делается обзор проблем методов генерации текстов, таких как несоответствие стилю и смыслу оригинала, потеря связности и логики, низкая уникальность и т.д.

В третьей главе описывается создание моделей машинного обучения для генерации текста на основе литературных произведений. В качестве основных моделей машинного обучения для генерации текстов были выбраны Long short-term memory model, Generative Pre-trained Transformer 2 и модель, основанная на цепях Маркова. LSTM был выбран как наиболее популярный и зарекомендовавший себя метод, способный генерировать качественные тексты. Его достаточно просто реализовать, однако каждый раз нужно обучать модель заново, что несколько замедляет получение результатов. В качестве второй модели было решено взять один из трансформеров, в частности Generative Pre-trained Transformer 2 (GPT2). Благодаря существующим наборам данных возможно пропустить процесс обучения, взяв уже предобученную модель, и просто дообучить её для работы с необходимыми данными. Марковская модель была выбрана, как давно зарекомендовавший себя способ генерации текстов. Его простота так же позволяет получать результаты в короткие сроки.

Обучение моделей Маркова и LSTM проводится на входных данных, предоставляемых пользователем. GPT2 имеет уже предобученную модель, поэтому для неё проводился только поиск удовлеворительных параметров генерации.

В качестве языка реализации был выбран Python, как наиболее подходящий для работы с нейросетями. Благодаря наличию множества библиотек

для машинного обучения, таких как PyTorch и Keras, этот язык программирования предоставляет лучшие средства для реализации подобных проектов. Платформой для реализации будет выступать Google Colab – бесплатная платформа для блокнотов Jupyter. Кроме среды запуска блокнотов Python и R, Colab позволяет совместно использовать свободный доступ к ограниченному количеству GPU и TPU, что являлось решающим фактором при выборе, так как при обучении трансформера понадобится использование технологии CUDA для увеличения мощности параллельных вычислений.

Также создается веб-интерфейс для реализованных моделей, который позволяет пользователю выбирать литературное произведение, метод генерации и длину текста. В конце проводится сравнение результатов работы алгоритмов.

ЗАКЛЮЧЕНИЕ

В дипломной работе были рассмотрены существующие методы анализа и генерации текста на основе литературных произведений, а также созданы и сравнены модели машинного обучения для генерации текста. В работе были выявлены основные проблемы и ограничения при анализе и генерации текста, такие как несоответствие стилю и смыслу оригинала, потеря связности и логики, низкая уникальность и т.д. Для решения этих проблем были предложены различные подходы, включая использование марковских цепей, LSTM модели и модели на основе GPT2.

В результате работы были созданы модели генерации текста на основе LSTM, GPT2 и Марковских цепей. Также был создан веб-интерфейс для реализованных моделей, который позволяет пользователю генерировать текст на основе введённого текста. Было проведено сравнение результатов генерации. В целом, работа показала, что генерация текста на основе литературных произведений является интересным и перспективным направлением исследований, которое может иметь множество практических применений в разных сферах.

Основные источники информации:

1 Fu, Zihao. A theoretical analysis of the repetition problem in text generation. — 2021

2 Szymanski, Grzegorz & Ciota, Zygmunt. Hidden Markov Models Suitable for Text Generation. — 2023

3 Wang, Ke. Sentigan. Generating sentimental texts via mixture adversarial networks. — 2018. — Pp. 4446–4452.

4 Tianyu Liu, Kexiang Wang, Lei Sha, Baobao Chang, Zhifang Sui. Table-to-text Generation by Structure-aware Seq2seq Learning. — 2017

5 Hochreiter, Sepp. Long short-term memory / Sepp Hochreiter, Jurgen Schmidhuber *Neural computation.* — 12 1997. — Vol. 9. — Pp. 1735–80.

6 Greff, Klaus. LSTM: A search space odyssey / Klaus Greff, Rupesh K. Srivastava, Jan Koutnik, Bas R. Steunebrink, Jurgen Schmidhuber // *IEEE Transactions on Neural Networks and Learning Systems.* — oct 2017. — Vol. 28, no. 10. — Pp. 2222–2232.

7 Кривошеев, Н.А. Генерация текста на основе нейронной сети lstm / Н.А. Кривошеев, К.В. Вик, Ю.А. Иванова, В.Г. Спицын // *Сборник*

трудов по материалам VII Международной конференции и молодежной школы. Том 3. — 2021.

8 Li, Yang. A generative model for category text generation / Yang Li, Quan Pan, Suhang Wang, Tao Yang, Erik Cambria // *Information Sciences.* — 03 2018. — Vol. 450.

9 Raghu, Maithra. Do vision transformers see like convolutional neural networks? — 2022.

10 Sun, Yu. Ernie 2.0: A continual pre-training framework for language understanding. 2019.