Аспекты создания корпоративной вопросно-ответной системы с использованием генеративных предобученных языковых моделей
Голиков Алексей Александрович аспирант, Отделение филологии и литературы, Кафедра русского языка и литературы, Казанский (Приволжский) федеральный университет (Елабужский институт) 109316, Россия, Москва, г. Москва, ул. Волгоградский Пр., 42 Golikov Aleksei Postgraduate student, Department of Philology and Literature, Department of Russian Language and Literature, Kazan (Volga Region) Federal University (Yelabuga Institute) 109316, Russia, Moscow, Volgogradsky Ave., 42 ag@mastercr.ru Другие публикации этого автора Акимов Дмитрий Андреевич ORCID: 0009-0004-2800-4430 кандидат технических наук Аналитик, ООО "Мастерская цифровых решений" 109316, Россия, Москва, г. Москва, Волгоградский пр., 42 Akimov Dmitrii PhD in Technical Science Analyst, LLC "Digital solutions workshop" 109316, Russia, Moscow, Volgogradsky ave., 42 akimovdmitry1@mail.ru Другие публикации этого автора Романовский Максим Сергеевич Sr. Technology Manager, Deutsche Bank AG 10243, Германия, Берлин, г. Берлин, ул. Koppenstraße, 93 Romanovskii Maksim Sr. Technology Manager, Deutsche Bank AG 10243, Germany, Berlin, Koppenstra straße, 93 maksim.s.romanovskii@gmail.com Тращенков Сергей Викторович ORCID: 0000-0001-8786-8336 Заведующий кафедрой программирования и вычислительных технологий Академии цифрового образования, ООО «Мобильное электронное образование» 127018, Россия, Москва, г. Москва, Сущёвский Вал, 16, стр. 4 Trashchenkov Sergei Head of the Department of Programming and Computing Technologies of the Academy of Digital Education, LLC «Mobile e-Learning» 127018, Russia, Moscow, Sushchevsky Val, 16, p. 4 trashchenkov@gmail.com
DOI: 10.25136/2409-8698.2023.12.69353 EDN: FSTHRW Дата направления статьи в редакцию: 17-12-2023 Дата публикации: 25-12-2023
Аннотация: В статье описаны различные способы использования генеративных предобученных языковых моделей для построения корпоративной вопросно-ответной системы. Существенным ограничением текущих генеративных предобученных языковых моделей является лимит по числу входных токенов, не позволяющий им работать «из коробки» с большим количеством документов или с документом большого размера. Для преодоления данного ограничения в работе рассмотрена индексация документов с последующим поисковым запросом и генерацией ответа на базе двух наиболее популярных на текущий момент open source решений – фреймворков Haystack, LlamaIndex. Было показано, что применение open source фреймворка Haystack при лучших настройках позволяет получить более точные ответы при построении корпоративной вопросно-ответной системы по сравнению с open source фреймворком LlamaIndex, однако требует использования в среднем несколько бóльшего числа токенов. В

статье использовался сравнительный анализ для оценки эффективности использования генеративных предобученных языковых моделей в корпоративных вопросно-ответных системах с помощью фреймворков Haystack и Llamaindex. Оценка полученных результатов осуществлялась с использованием метрики EM (exact match). Основными выводами проведенного исследования по созданию вопросно-ответных систем с использованием генеративных предобученных языковых моделей являются: 1. Использование иерархической индексации на текущий момент чрезвычайно затратно с точки зрения числа используемых токенов (около 160000 токенов для иерархической индексации против 30000 токенов в среднем для последовательной индексации), поскольку ответ генерируется путем последовательной обработки родительских и дочерних узлов. 2. Обработка информации при помощи фреймворка Haystack при лучших настройках позволяет получить несколько бóльшую точность ответов, чем использование фреймворка LlamaIndex (0.7 против 0.67 при лучших настройках). 3. Использование фреймворка Haystack более инвариантно относительно точности ответов с точки зрения количества токенов в чанке. 4. В среднем использование фреймворка Haystack более затратно по числу токенов (примерно в 4 раза), чем фреймворка LlamaIndex. 5. Режимы генерации ответа «create and refine» и «tree summarize» для фреймворка LlamaIndex являются примерно одинаковыми с точки зрения точности получаемых ответов, однако для режима «tree summarize» требуется больше токенов.
Ключевые слова: генеративные языковые модели, информационная поисковая система, вопросно-ответная система, индексация, Haystack, LlamaIndex, чанк, точность, токен, ретривер
Abstract: The article describes various ways to use generative pre-trained language models to build a corporate question-and-answer system. A significant limitation of the current generative pre-trained language models is the limit on the number of input tokens, which does not allow them to work "out of the box" with a large number of documents or with a large document. To overcome this limitation, the paper considers the indexing of documents with subsequent search query and response generation based on two of the most popular open source solutions at the moment – the Haystack and LlamaIndex frameworks. It has been shown that using the open source Haystack framework with the best settings allows you to get more accurate answers when building a corporate question-and-answer system compared to the open source LlamaIndex framework, however, requires the use of an average of several more tokens. The article used a comparative analysis to evaluate the effectiveness of using generative pre-trained language models in corporate question-and-answer systems using the Haystack and Llamaindex frameworks. The evaluation of the obtained results was carried out using the EM (exact match) metric. The main conclusions of the conducted research on

the creation of question-answer systems using generative pre-trained language models are: 1. Using hierarchical indexing is currently extremely expensive in terms of the number of tokens used (about 160,000 tokens for hierarchical indexing versus 30,000 tokens on average for sequential indexing), since the response is generated by sequentially processing parent and child nodes. 2. Processing information using the Haystack framework with the best settings allows you to get somewhat more accurate answers than using the LlamaIndex framework (0.7 vs. 0.67 with the best settings). 3. Using the Haystack framework is more invariant with respect to the accuracy of responses in terms of the number of tokens in the chunk. 4. On average, using the Haystack framework is more expensive in terms of the number of tokens (about 4 times) than the LlamaIndex framework. 5. The "create and refine" and "tree summarize" response generation modes for the LlamaIndex framework are approximately the same in terms of the accuracy of the responses received, however, more tokens are required for the "tree summarize" mode. Keywords: generative anguage models, information retrieval system, QA-system, indexing, Haystack, LlamaIndex, chunk, exact match, token, retriever 1 Введение Вопросно-ответные системы появились в 1960-х годах [1], и, как и другие области компьютерной лингвистики, с развитием технологий машинного обучения в последние годы претерпели существенные изменения. Вопросно-ответные системы бывают двух видов – экстрактивные и генеративные [2]. Экстрактивные вопросно-ответные системы в общем случае в качестве ответа выдают короткий ответ на заданный вопрос, часто в виде цитаты из поданного на вход набора документов. Например, на вопрос «в каком году родился лорд Байрон?» подобная система может ответить «в 1788 году», если данной системе были поданы для обработки материалы, содержащие биографию лорда Байрона, которые, скорее всего, включали в себя предложение «Лорд Байрон родился в 1788 году». Стоит отметить, что аналогичной функцией обладают и многие поисковые системы: например, «Google» выдаст подобный точный ответ выше различных ссылок на различные сайты. Интерес к генеративным языковым моделям (и генеративным вопросно-ответным системам, в частности) резко возрос после появления больших предобученных моделей GPT-3 и ChatGPT [3], впечатляющих своей «эрудицией» и способностью к сложным рассуждениям. Генеративные вопросно-ответные системы позволяют отвечать более развернуто на более сложные вопросы. Так, на вопрос «что общего у Лермонтова и Байрона?» экстрактивная вопросно-ответная система, скорее всего, не сможет дать ответ, если в поданных ей материалах не приведено подобное сравнение, в то время как модель ChatGPT (которую в данном контексте можно считать генеративной вопросно-ответной системой) дает развернутый ответ на заданный вопрос: «оба были поэтами-романтиками, имели репутацию

бунтарей, в какой-то момент были изгнанниками, известны своим лиризмом и умением передать красоту и природы и т.д.». На текущий момент (декабрь 2023 г.) ChatGPT-3.5 (бесплатная версия ChatGPT) была обучена на огромном числе общедоступных материалов, существовавших в мире до января 2022 года, поэтому они способны отвечать на большое количество вопросов по тем или иным отраслям знаний и по умолчанию неспособны отвечать на вопросы по событиям после января 2022 года или по тем данным, которые им не предоставлялись для обучения. При этом дополнительным существенным преимуществом была бы возможность подать данным моделям на вход собственные данные – будь то некая корпоративная документация, финансовые отчеты или новые научные статьи – чтобы иметь возможность получать ответы и рассуждения на их основе. Однако у столь мощных и привлекательных языковых моделей, как GPT-3 и ChatGPT, существует ограничение по числу токенов, которые могут быть поданы им на вход – так, например, подвид модели GPT-3 text-davinci-003 имеет ограничение в 4000 токенов на вход, т.е. около 3000 слов на английском языке. Таким образом, напрямую подать языковым моделям GPT-3 и ChatGPT на вход большое количество документов или документ большого размера (содержащий более 4000 токенов), чтобы получить ответы на вопросы по ним, невозможно. Другим подходом является дообучение модели на собственных дополнительных данных – что, однако, не всегда возможно как с технической точки зрения, так как требует значительных вычислительных ресурсов, так и с организационной, поскольку требует наличия высококлассных специалистов по анализу данных в штате компании. Третьим возможным подходом является суммаризация текстовых данных тем или иным способом до объема менее 4000 токенов, однако, очевидно, что значительная часть информации в таком случае будет потеряна. Во многих случаях наиболее привлекательным способом решения вопроса является индексация документов с последующим поисковым запросом и генерацией ответа, что возможно выполнить как полностью самостоятельно, так и с использованием популярных open source (с открытым исходным кодом) фреймворков Haystack, LlamaIndex, о чем и пойдет речь в данной статье. Стоит отметить, что исследованиям в области вопросно-ответных систем посвящено множество публикаций, однако в большинстве своем они посвящены экстрактивным вопросно-ответным системам, поскольку достаточно качественные генеративные большие языковые модели появились позже. Из наиболее свежих и релевантных научных материалов, касающихся в том числе генеративных вопросно-ответных систем, можно выделить статьи [4-6] и диссертацию [7]. Автор вышеупомянутой диссертации даже создал отдельный сервис (https://demo.caire.ust.hk/), работающий как

генеративная вопросно-ответная система по большому количеству статей о коронавирусе. Однако, отдавая должное автору вышеупомянутых диссертации и сервиса, стоит сказать, что в настоящее время, в частности, за счет появившихся уже после релиза ChatGPT фреймворков LlamaIndex, а также возможности использования фреймворка Haystack совместно с GPT-3 для построения генеративной вопросно-ответной системы, создание подобной системы стало значительно проще и доступнее, а потому сравнение фреймворков и их настроек актуально и представляет значительный интерес. 2 Расширение возможностей применения больших языковых моделей путем использования индексации документов Основным способом построения вопросно-ответных систем является использование ретривера для определения наиболее релевантных запросу частей текста, а затем синтез ответа из найденных частей текста с использованием т.н. ридера (для экстрактивной вопросно-ответной системы) или генератора (для генеративной вопросно-ответной системы). При этом для более эффективного поиска целесообразно предварительно провести т.н. индексацию документа или набора документов, по которым предполагается производить поиск. Под индексацией подразумевается выявление и сохранение некоей ключевой информации о частях документов, с помощью которой в дальнейшем удобно определить, насколько та или иная часть текста соответствует поисковому запросу (рисунок 1). Рисунок 1: Упрощенная схема работы вопросно-ответной системы с использованием индексации В качестве простого примера можно привести индексацию с использованием ключевых слов: для каждой части текста сохраняются ключевые термины, о которых идет речь, и далее при поисковом запросе будет выполнено сопоставление терминов запроса и сохраненных ключевых слов участков текста. Так, при поиске по странице «Википедии», посвященной лорду Байрону, при поисковом запросе «В каком году родился лорд Байрон?» по ключевым словам «рождение», «родиться» ретривером может быть найден участок текста «Джордж Гордон Байрон родился 22 января 1788 года». И далее задача ридера будет заключаться в извлечении из найденного участка текста требуемой информации – т.е. «1788 год» в данном случае. Очевидно, что подобный пример индексации с использованием ключевых слов хотя и несложен, но в то же время не слишком эффективен, поскольку в данном случае не совсем ясно, какие слова считать ключевыми. Одним из более предпочтительных в большинстве случаев способов индексации является индексация с использованием статистической меры TF-IDF, отражающей важность слова в корпусе, или же вариации TF-IDF – алгоритма BM25 [8, 9]. Так, при использовании TF-IDF в приведенном выше вопросе словам «году», «родился», «лорд», «Байрон»

автоматически будет назначен бóльший вес при поиске, так как они встречаются реже, чем слова «в» и «каком». Однако и в данном случае имеет место существенный недостаток, связанный с тем, что данные способы игнорируют порядок слов, контекст, возможность замены слова синонимами и т.д. С изобретением векторных семантических моделей появилась возможность производить индексацию документа, соотнося участку текста определенное представление в векторном пространстве, т.н. эмбеддинг [10, 11]. Данный способ индексации позволяет определять смысловой контекст, преодолевая таким образом недостатки способов индексации с использованием ключевых слов, TF-IDF и BM25. С появлением больших языковых моделей, таких как BERT, GPT и их вариаций, появилась возможность построения достаточно точных эмбеддингов в векторном пространстве большой размерности. В работе был выбран способ построения эмбеддингов с использованием подвида модели GPT-3 – ada-002 от Open AI [12] – входному тексту сопоставляется вектор в пространстве размерностью 1536. Также помимо выбора модели для построения индекса может быть выбран способ построения индекса – набор последовательных эмбеддингов, соответствующих последовательным частям текста (vector store index [13]) (рисунок 2) или же иерархическая древовидная структура индекса индексов (tree index) (рисунок 3), заключающаяся в последовательной восходящей суммаризации частей текста. Рисунок 2: Последовательная индексация Рисунок 3: Иерархическая индексация Также в фреймворке LlamaIndex возможны два режима генерации ответа на базе отобранных релевантных частей текста – итерационное улучшение ответа на базе каждой следующей релевантной части текста (режим «create and refine») (рисунок 4) и иерархическая суммаризация ответа на базе релевантных частей текста (режим «tree summarize») (рисунок 5). Рисунок 4: Итерационное улучшение ответа Рисунок 5: Иерархическая суммаризация ответа Стоит отметить, что второй популярный фреймворк для индексации документов – Haystack – не позволяет настолько гибко выбирать способы индексации и режимы генерации ответа. Фреймворк Haystack по умолчанию использует последовательную индексацию. Оба фреймворка способны обрабатывать различные оптимизированные хранилища векторных представлений, такие как Weaviate, Pinecone, FAISS и прочие [14, 15]. 3 Результаты испытания вопросно-ответной системы Для оценки качества работы описанных выше способов обработки текстовых данных достаточно большого объема был выбран документ от апреля 2022 года – т.н. «Белая книга искусственного интеллекта» от Китайской академии информационно-коммуникационных технологий в переводе на английский язык (https://cset.georgetown.edu/wp-content/uploads/t0442_AI_white_paper_2022_EN.pdf), содержащая около

12 тысяч слов (что составляет около 16 тысячи токенов), что примерно в 4 раза больше лимита на обработку текстовых данных за один раз текущей моделью GPT-3. Для данного документа был составлен вручную датасет вопросов и ответов (ответы человека принимались за эталонные). Для тестирования использовались open-source фреймворки Haystack и LlamaIndex, исходный текст для различных сценариев тестирования был разбит на части (т.н. чанки) по 20, 100, 200 и 1000 токенов. Для того, чтобы при делении исходного текста на части потеря информации была минимальной, использовалось наложение одной части текста на соседние части на 3 токена. Таким образом, полные параметры испытания вопросно-ответной системы с использованием генеративных предобученных языковых моделей следующие (Таблица 1). Параметр Значение Фреймворк Haystack; LlamaIndex Ретривер ada-002 (подвид модели GPT-3) Генератор davinci-003 (подвид модели GPT-3) Структура индекса (только для LlamaIndex) последовательная; иерархическая Режим синтеза ответа (только для LlamaIndex) create and refine; tree summarize Размер частей текста 20; 100; 200; 1000 токенов Наложение частей текста друг на друга 3 токена Метрика Exact match Таблица 1: Параметры испытания вопросно-ответной системы При тестировании вопросно-ответной системы измерялось количество затраченных токенов, которые необходимы для оценки потенциальных расходов на использование моделей OpenAI. При этом, поскольку для создания эмбеддингов применялся очень недорогой для использования подвид модели GPT-3 ada-002, то подсчет токенов для построения эмбеддингов не производился. В работе учитывались только токены, задействованные при генерации ответа с использованием более дорогого подвида модели GPT-3 davinci-003. Стоит отметить, что при использовании иерархической древовидной структуры индекса (рисунок 3) на ответы вопросно-ответной системы было затрачено 167510 токенов, что значительно больше среднего расхода токенов на ответы без использования данной индексной структуры (около 30000 токенов). Поэтому далее от иерархической индексации пришлось отказаться, в рамках данного исследования использовалась только последовательная индексация. В таблице 2 приведен пример ответа вопросно-ответной системы при различных настройках на один из вопросов, заданных по выбранному документу: «By 2025, will there be more data in the world – structured or unstructured – and by how many times?». Номер модели Параметры вопросно-ответной системы Ответ Верно / неверно 1 Фреймворк: Haystack Чанк: 20 токенов By 2025, there will be 163ZB of data in the world, with unstructured data accounting for the majority of the data, estimated to be 80%-90%. Да 2 Фреймворк: Haystack Чанк: 100 токенов By 2025, there will be more unstructured data, estimated to be 80%-90%, with the global data scale estimated to be 163ZB. Да 3 Фреймворк: Haystack

Чанк: 200 токенов By 2025, the global data scale is estimated to reach 163ZB, with unstructured data accounting for 80%-90%. Да 4 Фреймворк: Haystack Чанк: 1000 токенов By 2025, the global data scale will reach 163ZB, with unstructured data accounting for 80%-90%. Да 5 Фреймворк: LlamaIndex Индексация: последовательная Чанк: 20 токенов Режим генерации ответа: create and refine By 2025, it is likely that there will be more unstructured data in the world, potentially up to five times more than there is today. This is due to the fact that unstructured data currently accounts for 80% of the world's data, and this percentage is expected to increase as more data is generated. Нет 6 Фреймворк: LlamaIndex Индексация: последовательная Чанк: 20 токенов Режим генерации ответа: tree summarize By 2025, there will likely be more unstructured data in the world, potentially up to five times more than there is today. Нет 7 Фреймворк: LlamaIndex Индексация: последовательная Чанк: 100 токенов Режим генерации ответа: create and refine By 2025, there will be more unstructured data in the world, by a ratio of 80%-90%. Да 8 Фреймворк: LlamaIndex Индексация: последовательная Чанк: 100 токенов Режим генерации ответа: tree summarize By 2025, there will be more unstructured data in the world, by a ratio of 80%-90%. This means that there will be at least 80% more unstructured data than structured data in the world by 2025. Да 9 Фреймворк: LlamaIndex Индексация: последовательная Чанк: 200 токенов Режим генерации ответа: create and refine By 2025, there will be more unstructured data in the world, by a factor of 8-9 times. Да 10 Фреймворк: LlamaIndex Индексация: последовательная Чанк: 200 токенов Режим генерации ответа: tree summarize By 2025, there will be more unstructured data in the world, estimated to be 80%-90% of the total data. This is an increase of approximately 163ZB, which is approximately 163 times more data than there is currently. Да 11 Фреймворк: LlamaIndex Индексация: последовательная Чанк: 1000 токенов Режим генерации ответа: create and refine It is difficult to predict the amount of data in the world by 2025. However, it is likely that the amount of data will increase significantly, both in terms of structured and unstructured data. The growth of data is expected to be driven by the increasing use of artificial intelligence and machine learning technologies, as well as the increasing availability of data sources. Нет 12 Фреймворк: LlamaIndex Индексация: последовательная Чанк: 1000 токенов Режим генерации ответа: tree summarize It is impossible to predict by how many times the amount of data will increase by 2025, but it is likely that there will be more data in the world - both structured and unstructured - by 2025. Нет Таблица 2: Пример ответа вопросно-ответной системы на вопрос «By 2025, will there be more data in the world – structured or unstructured – and by how many times?» при различных настройках Для вопросно-ответной системы при различных настройках из таблицы 2 была определена доля правильных ответов (рисунок 6) – метрика EM (Exact match) [16, 17]: где

M – количество правильных ответов, N – общее количество вопросов в наборе данных для оценки. Рисунок 6: Доля правильных ответов при различных настройках вопросно-ответной системы (синий цвет – фреймворк Haystack, оранжевый – LlamaIndex, отсутствие штриховки – режим генерации ответа «create and refine», штриховка – режим генерации ответа «tree summarize») Также было определено количество токенов, затраченных при использовании подвида модели GPT-3 – davinci-003 – для генерации ответов на вопросы. Рисунок 7: Количество затраченных токенов ответов при различных настройках вопросно-ответной системы (синий цвет – фреймворк Haystack, оранжевый – LlamaIndex, отсутствие штриховки – режим генерации ответа «create and refine», штриховка – режим генерации ответа «tree summarize») Таким образом, наибольшую точность ответов продемонстрировала вопросно-ответная система, использующая open-source фреймворк Haystack, при количестве токенов в чанке 100, 200, 1000 (для всех трех случаев точность одинаковая и составляет 0.7). При этом, как видим из рисунка 7, чем больше токенов в чанке, тем больше требуется использовать токенов при генерации ответа – что логично, поскольку генератор создает ответ, обрабатывая отобранные ретривером чанки, которые тем больше по размеру, чем больше токенов в чанке. Для фреймворка LlamaIndex режимы генерации ответа «create and refine» и «tree summarize» являются примерно одинаковыми с точки зрения точности получаемых ответов, однако для режима «tree summarize» требуется больше токенов. 4 Заключение Генеративные предобученные языковые модели (такие как ChatGPT) произвели революцию в области обработки естественного языка. Однако их существенным ограничением является их лимит по числу входных токенов, который может быть преодолен путем использования индексных структур данных. В работе было рассмотрено создание вопросно-ответной системы с использованием генеративных предобученных языковых моделей на базе двух основных open source фреймворков – Haystack и LlamaIndex. На базе документа «Белая книга искусственного интеллекта» от Китайской академии информационно-коммуникационных технологий, был составлен датасет вопросов и ответов для оценки качества работы вопросно-ответной системы при различных настройках с использованием метрики Exact match. В качестве результатов проведенного исследования можно привести следующие положения: 1. Использование иерархической индексации на текущий момент чрезвычайно затратно с точки зрения числа используемых токенов (около 160000 токенов для иерархической индексации против 30000 токенов в среднем для последовательной индексации), поскольку ответ генерируется путем последовательной обработки родительских и дочерних узлов. 2. Обработка информации при помощи фреймворка Haystack при лучших настройках позволяет

получить несколько бóльшую точность ответов, чем использование фреймворка LlamaIndex (0.7 против 0.67 при лучших настройках). 3. Использование фреймворка Haystack более инвариантно относительно точности ответов с точки зрения количества токенов в чанке – для количества токенов в чанке 100, 200 и 1000 точность ответов была одинакова и составила 0.7. 4. В среднем использование фреймворка Haystack более затратно по числу токенов (примерно в 4 раза), чем фреймворка LlamaIndex. 5. Режимы генерации ответа «create and refine» и «tree summarize» для фреймворка LlamaIndex являются примерно одинаковыми с точки зрения точности получаемых ответов, однако для режима «tree summarize» требуется больше токенов. Таким образом, применение open source фреймворка Haystack при лучших настройках позволяет получить более точные ответы при построении корпоративной вопросно-ответной системы по сравнению с open source фреймворком LlamaIndex, однако требует использования в среднем несколько бóльшего числа токенов. Библиография 1. Simmons R. F., Klein S., McConlogue K. Indexing and dependency logic for answering English questions // American Documentation. – 1964. – Т. 15. – №. 3. – С. 196-204. 2. Luo M. et al. Choose your qa model wisely: A systematic study of generative and extractive readers for question answering // arXiv preprint arXiv:2203.07522. – 2022. 3. Zhou C. et al. A comprehensive survey on pretrained foundation models: A history from bert to chatgpt // arXiv preprint arXiv:2302.09419. – 2023. 4. Lewis P. et al. Retrieval-augmented generation for knowledge-intensive nlp tasks //Advances in Neural Information Processing Systems. – 2020. – Т. 33. – С. 9459-9474. 5. Маслюхин С. М. Диалоговая система на основе устных разговоров с доступом к неструктурированной базе знаний // Научно-технический вестник информационных технологий, механики и оптики. – 2023. – Т. 23. – №. 1. – С. 88-95. 6. Евсеев Д. А., Бурцев М. С. Использование графовых и текстовых баз знаний в диалоговом ассистенте DREAM // Труды Московского физико-технического института. – 2022. – Т. 14. – №. 3 (55). – С. 21-33. 7. Su D. Generative Long-form Question Answering: Relevance, Faithfulness and Succinctness //arXiv preprint arXiv:2211.08386. – 2022. 8. Kim M. Y. et al. Legal information retrieval and entailment based on bm25, transformer and semantic thesaurus methods // The Review of Socionetwork Strategies. – 2022. – Т. 16. – №. 1. – С. 157-174. 9. Ke W. Alternatives to Classic BM25-IDF based on a New Information Theoretical Framework //2022 IEEE International Conference on Big Data (Big Data). – IEEE, 2022. – С. 36-44. 10. Rodriguez P. L., Spirling A. Word embeddings: What works, what doesn’t, and how to tell the difference for applied research // The Journal of Politics. – 2022. – Т. 84. – №. 1. – С. 101-115. 11. Жеребцова Ю. А., Чижик А. В. Сравнение моделей векторного представления текстов в задаче создания чат-бота // Вестник Новосибирского государственного

университета. Серия: Лингвистика и межкультурная коммуникация. – 2020. – Т. 18. – №. 3. – С. 16-34. 12. Digutsch J., Kosinski M. Overlap in meaning is a stronger predictor of semantic activation in GPT-3 than in humans //Scientific Reports. – 2023. – Т. 13. – №. 1. – С. 5035. 13. Kamnis S. Generative pre-trained transformers (GPT) for surface engineering // Surface and Coatings Technology. – 2023. – С. 129680. 14. Khadija M. A., Aziz A., Nurharjadmo W. Automating Information Retrieval from Faculty Guidelines: Designing a PDF-Driven Chatbot powered by OpenAI ChatGPT // 2023 International Conference on Computer, Control, Informatics and its Applications (IC3INA). – IEEE, 2023. – С. 394-399. 15. Johnson J., Douze M., Jégou H. Billion-scale similarity search with gpus // IEEE Transactions on Big Data. – 2019. – Т. 7. – №. 3. – С. 535-547. 16. Rajpurkar P. et al. Squad: 100,000+ questions for machine comprehension of text // arXiv preprint arXiv:1606.05250. – 2016. 17. Bai Y., Wang D. Z. More than reading comprehension: A survey on datasets and metrics of textual question answering // arXiv preprint arXiv:2109.12264. – 2021. References
1. Simmons, R. F., Klein, S., & McConlogue, K. (1964). Indexing and dependency logic for answering English questions. American Documentation, 15(3), 196-204. 2. Luo, M., Hashimoto, K., Yavuz, S., Liu, Z., Baral, C., & Zhou, Y. (2022). Choose your qa model wisely: A systematic study of generative and extractive readers for question answering. arXiv preprint arXiv:2203.07522. 3. Zhou, C., Li, Q., Li, C., Yu, J., Liu, Y., Wang, G., ... & Sun, L. (2023). A comprehensive survey on pretrained foundation models: A history from bert to chatgpt. arXiv preprint arXiv:2302.09419. 4. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., ... & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. Advances in Neural Information Processing Systems, 33, 9459-9474.
5. Maslyuhin, S. M. (2023). A spoken dialogue-based system with access to an unstructured knowledge base. Scientific and Technical Bulletin of Information Technologies, Mechanics and Optics, 23(1), 88-95 6. Evseev, D. A., Burcev, M. S. (2022). Use of graph and text knowledge bases in the dialogue assistant DREAM. Proceedings of the Moscow Institute of Physics and Technology, 14(3), 21-33. 7. Su, D. (2022). Generative Long-form Question Answering: Relevance, Faithfulness and Succinctness. arXiv preprint arXiv:2211.08386. 8. Kim, M. Y., Rabelo, J., Okeke, K., & Goebel, R. (2022). Legal information retrieval and entailment based on bm25, transformer and semantic thesaurus methods. The Review of Socionetwork Strategies, 16(1), 157-174. 9. Ke, W. (2022, December). Alternatives to Classic BM25-IDF based on a New Information Theoretical Framework. In 2022 IEEE International Conference on Big Data (Big Data) (pp. 36-44). IEEE.
10. Rodriguez, P. L., & Spirling, A. (2022). Word embeddings: What works, what doesn’t, and how to tell the difference for applied research. The Journal of Politics, 84(1), 101-115. 11. Zherebcova, Yu. A., Chizhik, A. V. (2020).

Comparison of text vector representation models in the task of chatbot creation. Bulletin of Novosibirsk State University. Series: Linguistics and Intercultural Communication, 18(3), 16-34. 12. Digutsch, J., & Kosinski, M. (2023). Overlap in meaning is a stronger predictor of semantic activation in GPT-3 than in humans. Scientific Reports, 13(1), 5035. 13. Kamnis, S. (2023). Generative pre-trained transformers (GPT) for surface engineering. Surface and Coatings Technology, 129680. 14. Khadija, M. A., Aziz, A., & Nurharjadmo, W. (2023, October). Automating Information Retrieval from Faculty Guidelines: Designing a PDF-Driven Chatbot powered by OpenAI ChatGPT. In 2023 International Conference on Computer, Control, Informatics and its Applications (IC3INA) (pp. 394-399). IEEE. 15. Johnson, J., Douze, M., & Jégou, H. (2019). Billion-scale similarity search with gpus. IEEE Transactions on Big Data, 7(3), 535-547. 16. Rajpurkar, P., Zhang, J., Lopyrev, K., & Liang, P. (2016). Squad: 100,000+ questions for machine comprehension of text. arXiv preprint arXiv:1606.05250. 17. Bai, Y., & Wang, D. Z. (2021). More than reading comprehension: A survey on datasets and metrics of textual question answering. arXiv preprint arXiv:2109.12264. Результаты процедуры рецензирования статьи В связи с политикой двойного слепого рецензирования личность рецензента не раскрывается. Со списком рецензентов издательства можно ознакомиться здесь. Тема рецензируемой статьи, безусловно, является актуальной; автор данного труда касается вопроса использования вопросно-ответной системы в рамках платформ генеративных предобученных языковых моделей. Как отмечается в начале исследования, «вопросно-ответные системы появились в 1960-х годах, и, как и другие области компьютерной лингвистики, с развитием технологий машинного обучения в последние годы претерпели существенные изменения. Вопросно-ответные системы бывают двух видов – экстрактивные и генеративные. Экстрактивные вопросно-ответные системы в общем случае в качестве ответа выдают короткий ответ на заданный вопрос, часто в виде цитаты из поданного на вход набора документов», «интерес к генеративным языковым моделям (и генеративным вопросно-ответным системам, в частности) резко возрос после появления больших предобученных моделей GPT-3 и ChatGPT [3], впечатляющих своей «эрудицией» и способностью к сложным рассуждениям». Статья грамотно структурирована, ее наличного объема достаточно для раскрытия темы, обозначения аргументационной базы, манифестации суждений / выводов / умозаключений. Автор подробно рассматривает вопросно-ответный механизм, который является основной генеративных предобученных языковых моделей, таких как ChatGPT, выверяет / систематизирует основной блок критических источников, оценивает продуктивность указанной формы. Стиль работы ориентирован на собственно научный тип; статья дифференцирована на

смысловые блоки, общая аналитическая логика выровнена на протяжении всего труда. Материал достаточно информативен: «основным способом построения вопросно-ответных систем является использование ретривера для определения наиболее релевантных запросу частей текста, а затем синтез ответа из найденных частей текста с использованием т.н. ридера (для экстрактивной вопросно-ответной системы) или генератора (для генеративной вопросно-ответной системы)». Цитатный пласт сопровождается комментарием; считаю, что работа может быть полезна при формировании новых исследований смежной тематической направленности. Практическая составляющая материала заключается в том, что «для тестирования использовались open-source фреймворки Haystack и LlamaIndex, исходный текст для различных сценариев тестирования был разбит на части (т.н. чанки) по 20, 100, 200 и 1000 токенов. Для того, чтобы при делении исходного текста на части потеря информации была минимальной, использовалось наложение одной части текста на соседние части на 3 токена», «при тестировании вопросно ответной системы измерялось количество затраченных токенов, которые необходимы для оценки потенциальных расходов на использование моделей OpenAI. При этом, поскольку для создания эмбеддингов применялся очень недорогой для использования подвид модели GPT-3 ada-002, то подсчет токенов для построения эмбеддингов не производился. В работе учитывались только токены, задействованные при генерации ответа с использованием более дорогого подвида модели GPT-3 davinci-003». Полученные в ходе анализа данные структурированы в табличный вид, сведение данных в единый блок оправдано. Стандарт оформления выдержан, необходимые пометы сделаны: например, «Рисунок 6: Доля правильных ответов при различных настройках вопросно-ответной системы (синий цвет – фреймворк Haystack, оранжевый – LlamaIndex, отсутствие штриховки – режим генерации ответа «create and refine», штриховка – режим генерации ответа «tree summarize»)» и т.д. Итоги работы сведены к следующему: «генеративные предобученные языковые модели (такие как ChatGPT) произвели революцию в области обработки естественного языка. Однако их существенным ограничением является их лимит по числу входных токенов, который может быть преодолен путем использования индексных структур данных. В работе было рассмотрено создание вопросно-ответной системы с использованием генеративных предобученных языковых моделей на базе двух основных open source фреймворков – Haystack и LlamaIndex. На базе документа «Белая книга искусственного интеллекта» от Китайской академии информационно-коммуникационных технологий, был составлен датасет вопросов и ответов для оценки качества работы вопросно-ответной системы при различных настройках с использованием метрики Exact match…»,

«применение open source фреймворка Haystack при лучших настройках позволяет получить более точные ответы при построении корпоративной вопросно-ответной системы по сравнению с open source фреймворком LlamaIndex, однако требует использования в среднем несколько бóльшего числа токенов». Список источников отражен в основном тексте, формат отсылки учтен. Считаю, что работа имеет полновесный вид, тема исследования раскрыта, материал может быть полезен заинтересованным читателям / исследователям указанной проблемы. Рекомендую статью «Аспекты создания корпоративной вопросно-ответной системы с использованием генеративных предобученных языковых моделей» к открытой публикации в научном журнале «Litera» ИД «Nota Bene».

Правильная ссылка на статью: Голиков А.А., Акимов Д.А., Романовский М.С., Тращенков С.В. Аспекты создания корпоративной вопросно-ответной системы с использованием генеративных предобученных языковых моделей // Litera. 2023. № 12. С. 190-205. DOI: 10.25136/2409-8698.2023.12.69353 EDN: FSTHRW URL: https://nbpublish.com/library_read_article.php?id=69353