РОССИЙСКАЯ ФЕДЕРАЦИЯ

(19) **RU** (11) **2 820 264**(13) **C1**

(51) МПК
G10L 15/18 (2013.01)
G06N 20/00 (2019.01)

ФЕДЕРАЛЬНАЯ СЛУЖБА
ПО ИНТЕЛЛЕКТУАЛЬНОЙ СОБСТВЕННОСТИ

(12) ОПИСАНИЕ ИЗОБРЕТЕНИЯ К ПАТЕНТУ

(52) СПК
G10L 15/18 (2024.01); G06N 20/00 (2024.01)

(21)(22) Заявка: **2023113361, 23.05.2023**

(24) Дата начала отсчета срока действия патента:
23.05.2023

Дата регистрации:
31.05.2024

Приоритет(ы):
(22) Дата подачи заявки: **23.05.2023**

(45) Опубликовано: **31.05.2024** Бюл. № **16**

Адрес для переписки:
119021, Москва, ул. Льва Толстого, 16, ООО "Яндекс", патентный отдел, Малахова Надия Ахавовна

(72) Автор(ы):
Зинов Николай Александрович (RU),
Коренев Артём Аркадьевич (RU)

(73) Патентообладатель(и):
ОБЩЕСТВО С ОГРАНИЧЕННОЙ ОТВЕТСТВЕННОСТЬЮ "ЯНДЕКС" (RU)

(56) Список документов, цитированных в отчете о поиске: **US 2020/0342354 A1, 29.10.2020. US 2022/0035998 A1, 03.02.2022. US 2017/0323636 A1, 09.11.2017. US 2023/0085061 A1, 16.03.2023. RU 2792405 C2, 21.03.2023.**

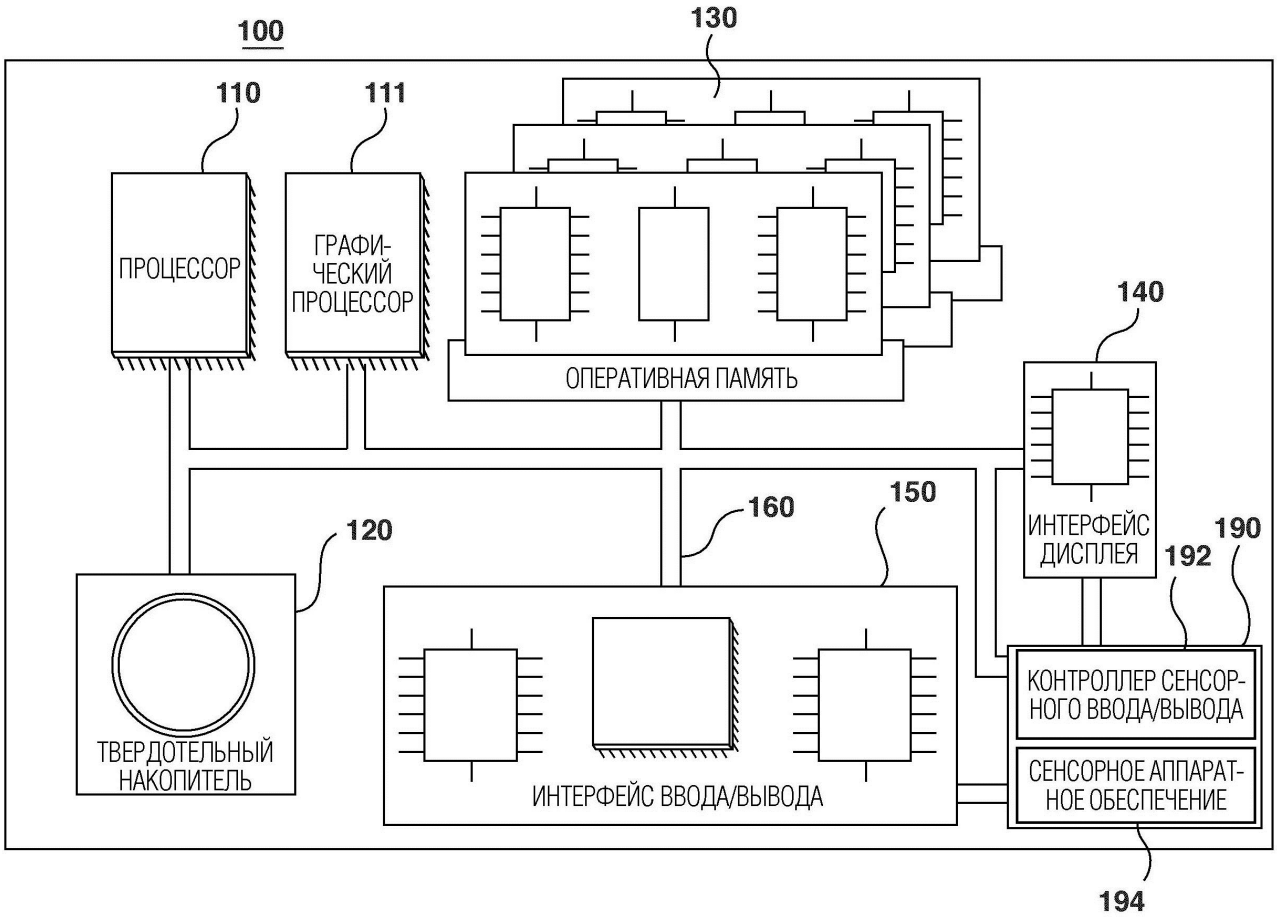
(54) СПОСОБ И СИСТЕМА ОБУЧЕНИЯ СИСТЕМЫ ЧАТ-БОТА

(57) Реферат:

Изобретение относится к области вычислительной техники. Технический результат заключается в обеспечении предоставления более естественных ответов на запросы пользователей, если наряду с заданным запросом пользователя система чат-бота может быть выполнена с возможностью учета факта, определенного как относящегося к заданному запросу пользователя. Технический результат достигается за счет получения диалоговых данных, включающих в себя текстовые представления диалоговых пар; получения данных о фактах, включающих в себя текстовые представления фактов; идентификации из данных о фактах для заданной диалоговой пары факта, относящегося к ней; создания обучающего набора данных, включающего в себя множество обучающих цифровых объектов, каждый из которых включает в себя текстовые представления: (i) человеческий запрос заданной диалоговой пары; (ii) факт; и (iii) человеческий ответ заданной диалоговой пары, отвечающий на человеческий запрос; подачи обучающего набора данных в систему чат-бота, заставляющего систему чат-бота формировать машинно-формируемый ответ; и оптимизации разницы между машинно-формируемым ответом и человеческим ответом, за счет чего происходит обучение системы чат-бота формировать машинно-формируемые ответы на запросы пользователей. 2 н. и 18 з.п. ф-лы, 8 ил.

RU 2820264 C1

RU 2820264 C1



ФИГ. 1

RU 2820264 C1

RU 2820264 C1

RUSSIAN FEDERATION

FEDERAL SERVICE
FOR INTELLECTUAL PROPERTY

(19) **RU** (11) **2 820 264**(13) **C1**

(51) Int. Cl.
G10L 15/18 (2013.01)
G06N 20/00 (2019.01)

(12) ABSTRACT OF INVENTION

(52) CPC
G10L 15/18 (2024.01); G06N 20/00 (2024.01)

(21)(22) Application: **2023113361, 23.05.2023**

(24) Effective date for property rights:
23.05.2023

Registration date:
31.05.2024

Priority:
(22) Date of filing: **23.05.2023**

(45) Date of publication: **31.05.2024** Bull. № **16**

Mail address:
119021, Moskva, ul. Lva Tolstogo, 16, OOO "Yandeks", patentnyj otdel, Malakhova Nadiya Akhavovna

(72) Inventor(s):
Zinov Nikolaj Aleksandrovich (RU),
Korenev Artem Arkadevich (RU)

(73) Proprietor(s):
OBSHCHESTVO S OGRANICHENNOJ OTVETSTVENNOSTYU "YANDEKS" (RU)

RU 2820264 C1

(54) METHOD AND SYSTEM FOR TRAINING CHATBOT SYSTEM

(57) Abstract:

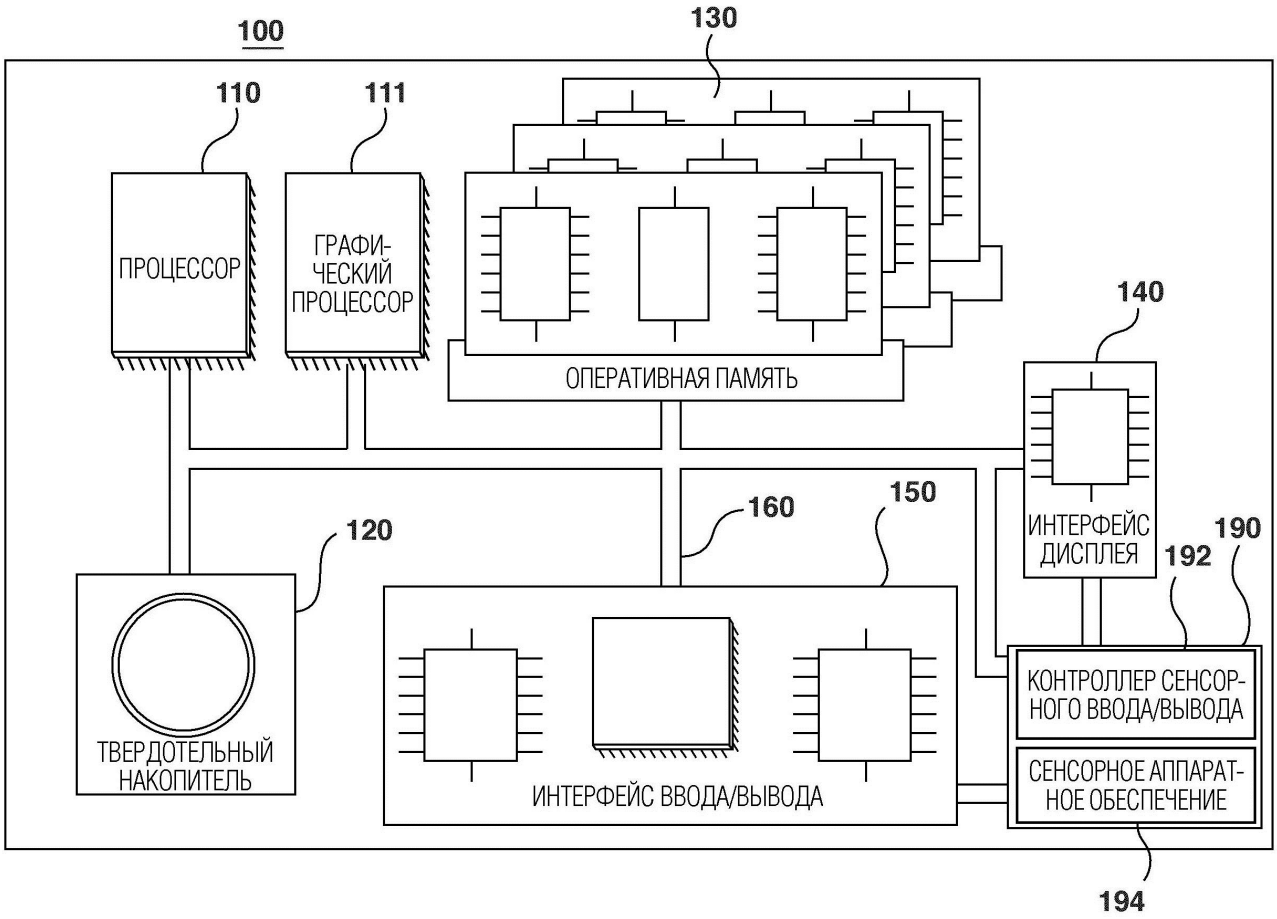
FIELD: computer engineering.

SUBSTANCE: technical result is achieved by obtaining dialogue data, including text representations of dialogue pairs; obtaining data on facts, including textual representations of facts; identification from data on facts for a given dialogue pair of fact relating to it; creating a training data set, which includes a plurality of training digital objects, each of which includes textual representations: (i) a human request of a given dialogue pair; (ii) fact; and (iii) a human response of a given dialogue pair, responding to a human request; feeding the training data set to the chatbot system forcing the chatbot system to generate a machine-generated response; and optimizing difference between machine-generated response and human response, due to which the chatbot system is trained to generate machine-generated responses to user requests.

EFFECT: providing more natural responses to user requests, if, along with a given user request, the chatbot system can be configured to take into account a fact determined as relating to a given user request.

20 cl, 8 dwg

RU 2820264 C1



ФИГ. 1

RU 2820264 C1

RU 2820264 C1

ОБЛАСТЬ ТЕХНИКИ

[001] Настоящая технология в целом относится к области обработки естественного языка; и, в частности, к способам и системам обучения системы чат-бота формировать ответы на естественном языке на запросы пользователей.

УРОВЕНЬ ТЕХНИКИ

[002] Системы чат-бота (также известные как «приложения виртуального помощника»), такие как система чат-бота Alexa™, система чат-бота ALISA™ или система чат-бота Siri™, могут обеспечивать улучшенное взаимодействие между пользователем и определенными онлайн-службами (например, платформами онлайн-покупок, системами онлайн-бронирования и т. п.) и/или их электронными устройствами. Это особенно актуально, когда пользователь является новичком или пользователем с ограниченными возможностями и, таким образом, в настоящее время не может использовать машинно-пользовательские интерфейсы таких онлайн-служб и/или интерфейсы электронного устройства для эффективного взаимодействия с ними. Например, пользователь, который находится за рулем, или пользователь с ослабленным зрением может быть не в состоянии использовать клавиатуру сенсорного экрана, связанную с их электронным устройством, для навигации по веб-сайту врача, чтобы записаться к нему на прием, или через платформу онлайн-покупок, чтобы отправить заказ. В то же время персонал службы поддержки клиентов этих онлайн-служб может быть недоступен для оказания помощи таким пользователям, например, из-за увеличения количества запросов от других пользователей.

[003] Например, данная система чат-бота может выполняться онлайн-платформой потокового видео (например, онлайн-платформой потокового видео Netflix™ или онлайн-платформой потокового видео Kinpoisk™) для помощи пользователям в выборе фильма или шоу для просмотра. В этом примере пользователь может отправить запрос на естественном языке, например: *«Привет, хочу посмотреть что-нибудь смешное»*. В ответ данная система чат-бота может быть настроена для (i) анализа запроса пользователя для определения намерений пользователя; (ii) передачи намерения пользователя, например, механизму рекомендаций онлайн-платформы потокового видео для формирования списка соответствующих комедийных шоу, соответствующих намерению пользователя; и (iii) предоставления ответа на запрос пользователя, включающего в себя список рекомендуемых комедийных шоу, например: *«Вот список комедийных шоу, которые вы искали: ...»*.

[004] В другом примере данная система чат-бота может выполняться платформой онлайн-листинга (такой как платформа онлайн-листинга Avito™ или платформа онлайн-листинга Kijiji™) для помощи пользователям в выборе товаров для покупки, доступных на платформе онлайн-листинга. В этом примере запрос пользователя может выглядеть так: *«Можете ли вы показать мне лучший телефон менее чем за 1000 долларов?»*. Точно так же в ответ система чат-бота может быть выполнена с возможностью формирования ответа, включающего в себя список телефонов, цены на которые ниже 1000 долларов США, определенные механизмом рекомендаций. Таким образом, данная система чат-бота может использоваться для навигации пользователя по соответствующей онлайн-службе и ответа на ее запросы.

[005] Однако одним из недостатков, связанных с такими системами чат-бота, является то, что предоставляемые ими ответы могут восприниматься как неестественные и безликие, не соответствующие стилю запросов пользователей. Это может повлиять на работу пользователей с данной системой чат-бота, заставляя пользователей избегать передачи своих запросов через системы чат-бота, что может привести к общей

неудовлетворенности пользователей соответствующими онлайн-службами.

[006] Были предложены определенные подходы предшествующего уровня техники для решения указанной выше технической проблемы.

[007] Статья за авторством Paranjape *и др.,* названная ***«HINDSIGHT: POSTERIOR-GUIDED TRAINING OF RETRIEVERS FOR IMPROVED OPEN-ENDED GENERATION»***, опубликованная Стэнфордским университетом 14 октября 2021 г., раскрывает использование дополнительного направляющего ретривера (англ. retreiver), которому разрешено использовать целевой вывод и «задним числом» извлекать соответствующие отрывки во время обучения. В частности, статья относится к ведущему ретриверу после апостериорного распределения Q отрывков с учетом входных и целевых выходных данных и обучении его совместно со стандартным ретривером и генератором путем максимизации нижней границы свидетельства (ELBo) в ожидании, превышающем Q. Для информативных бесед из набора данных Wizard of Wikipedia с апостериорным обучением ретривер находит отрывки с более высокой релевантностью в топ-10 (относительное улучшение 23%), ответы генератора более обоснованы в извлеченном отрывке (относительное улучшение 19%), а сквозная (англ. end-to-end) система обеспечивает более высокий общий результат (относительное улучшение 6,4%).

[008] **Патент США №: 11 068 660-B2** от 20 июля 2021 г., заявителя Koninklijke Philips NV, с названием ***«SYSTEMS AND METHODS FOR NEURAL CLINICAL PARAPHRASE GENERATION»***, раскрывает систему формирования парафраз, содержащую один или более аппаратных процессоров и/или других компонентов. Система выполнена с возможностью получения обучающего корпуса. Обучающего корпус включает в себя язык и известные парафразы языка. Система выполнена с возможность формирования на основе обучающего корпуса модели внимания на уровне слов и модели внимания на уровне символов. Система выполнена с возможностью предоставления одного или более парафраз-кандидатов ввода на естественном языке на основе моделей внимания как на уровне слов, так и на уровне символов. Модель, основанная на внимании на уровне слов, представляет собой двунаправленную сеть долговременной кратковременной памяти (LS™) на уровне слов, а модель, основанная на внимании на уровне символов, представляет собой двунаправленную сеть LS™ на уровне символов. Сети LS™ на уровне слов и символов создаются на основе слов и символов в обучающем корпусе. В некоторых вариантах осуществления сети LS™ представляют собой стекированные остаточные сети LS™, содержащие остаточные соединения между стекированными слоями данной сети LS™.

СУЩНОСТЬ ИЗОБРЕТЕНИЯ

[009] Целью настоящей технологии является устранение по меньшей мере некоторых неудобств, имеющихся в предшествующем уровне техники.

[0010] Разработчики настоящей технологии оценили, что данная система чат-бота может быть выполнена с возможностью предоставления более естественных ответов на запросы пользователей, если наряду с заданным запросом пользователя система чат-бота может быть выполнена с возможностью учета факта, определенного как относящегося к заданному запросу пользователя.

[0011] Более конкретно, в соответствии с некоторыми неограничивающими вариантами осуществления настоящей технологии система чат-бота может содержать: (i) модель машинного обучения семантического подобия, которая выполнена с возможностью формирования векторных вложений введенного в нее текста, например, текстовых представлений запросов пользователей; и (ii) генеративную модель, выполненную с возможностью формирования соответствующих ответов на запросы

пользователей на основе векторных вложений из модели семантического подобия.

[0012] Кроме того, по меньшей мере в некоторых неограничивающих вариантах осуществления настоящей технологии настоящие способы и системы направлены на обучение системы чат-бота формировать более естественные ответы на запросы пользователей путем подачи в генеративную модель, наряду с векторным вложением заданного обучающего пользовательского запроса, векторного вложения соответствующего факта, определенного как относящегося к данному обучающему пользовательскому запросу. Таким образом, система чат-бота может быть обучена предоставлять ответы на запросы пользователей в контексте соответствующих фактов, имеющих отношение к запросам пользователей, что, таким образом, может позволить улучшить стиль ответов, формируемых системой чат-бота. Другими словами, сформированные таким образом ответы системы чат-бота могут восприниматься пользователями как составляющие части естественной речи, аналогичной той, которую произносят люди. Это может повысить удовлетворенность пользователей системой чат-бота и связанной с ней онлайн-службой в целом.

[0013] Более конкретно, в соответствии с первым широким аспектом настоящей технологии предоставляется реализуемый компьютером способ обучения системы чат-бота формировать машинно-формируемые ответы на запросы пользователей системы чат-бота. Система чат-бота включает в себя: (i) модель машинного обучения (ML) семантического подобия, выполненную с возможностью идентификации соответствующих фактов, имеющих отношение к запросам пользователей; и (ii) генеративную ML-модель, подлежащую обучению для формирования текстовых представлений машинно-формируемых ответов на запросы пользователей на основе запросов пользователей и соответствующих фактов. Способ может выполняться сервером, включающим в себя процессор. Способ включает в себя: получение процессором диалоговых данных, включающих в себя (i) текстовые представления человеческих запросов диалогов пользователей на естественном языке; и (ii) текстовые представления соответствующих человеческих ответов диалогов, отвечающих на человеческие запросы; получение процессором данных о фактах, включающих в себя текстовые представления фактов; идентификацию процессором с использованием ML-модели семантического подобия из данных о фактах для заданной диалоговой пары, включающей в себя текстовые представления заданного человеческого запроса и соответствующий человеческий ответ, отвечающий на него, из диалоговых данных, текстового представления соответствующего факта, относящегося к по меньшей мере одному из заданного человеческого запроса и соответствующего человеческого ответа; формирование процессором обучающего набора данных, включающего в себя множество обучающих цифровых объектов, каждый из которых включает в себя: (i) текстовое представление заданного человеческого запроса; (ii) текстовое представление соответствующего факта; и (iii) соответствующую метку, являющуюся текстовым представлением соответствующего человеческого ответа на заданный человеческий запрос; подачу процессором обучающего набора данных в систему чат-бота, при этом подача включает в себя: для заданного обучающего цифрового объекта из множества обучающих цифровых объектов, подачу процессором объединения текстового представления заданного человеческого запроса и текстового представления соответствующего факта в генеративную ML-модель для формирования текстового представления соответствующего машинно-формируемого ответа на заданный человеческий запрос с учетом контекста соответствующего факта; и оптимизацию с помощью процессора разницы между соответствующим машинно-формируемым

ответом и соответствующим человеческим ответом на заданный человеческий запрос, за счет чего происходит обучение системы чат-бота формировать машинно-формируемые ответы на запросы пользователей.

[0014] В некоторых реализациях способа идентификация с использованием ML-модели семантического подобия соответствующего факта для заданной диалоговой пары текстовых представлений включает в себя: подачу процессором каждого из: (i) текстовых представлений человеческих запросов; (ii) текстовых представлений человеческих ответов; и (iii) текстовых представлений фактов в ML-модель семантического подобия для формирования соответствующих векторных вложений каждого из (i) текстовых представлений человеческих запросов; (ii) текстовых представлений человеческих ответов; и (iii) текстовых представлений фактов; отображение процессором соответствующих векторных вложений в векторное пространство; и идентификацию процессором соответствующего факта, относящегося к по меньшей мере одному из заданного человеческого запроса и соответствующего человеческого ответа, как факта, связанного с соответствующим векторным вложением, которое наиболее близко к соответствующему векторному вложению по меньшей мере одного из заданного человеческого запроса и соответствующего человеческого ответа в векторном пространстве.

[0015] В некоторых реализациях способа идентификация дополнительно содержит применение процессором алгоритма k ближайших соседей.

[0016] В некоторых реализациях способа идентификация дополнительно включает в себя применение процессором эвристического алгоритма.

[0017] В некоторых реализациях способа эвристический алгоритм содержит функцию ранжирования.

[0018] В некоторых реализациях способа функция ранжирования представляет собой функцию ранжирования BM25.

[0019] В некоторых реализациях способа соответствующий факт предназначен для предоставления контекста заданной диалоговой паре.

[0020] В некоторых реализациях способа объединение текстового представления заданного человеческого запроса и текстового представления соответствующего факта включает в себя объединение соответствующих векторных вложений заданного человеческого запроса и соответствующего факта в заданном векторном пространстве.

[0021] В некоторых реализациях способа оптимизация разницы включает в себя оптимизацию функции потерь, представляющей разницу между соответствующим машинно-формируемым ответом и соответствующим человеческим ответом.

[0022] В некоторых реализациях способа способ дополнительно включает в себя использование системы чат-бота для формирования машинно-формируемых ответов на запросы пользователей, при этом использование содержит: получение процессором текстового представления используемого человеческого запроса заданного пользователя; идентификацию процессором с использованием ML-модели семантического подобия из данных о фактах соответствующего факта, относящегося к используемому человеческому запросу; подачу процессором объединения текстовых представлений используемого человеческого запроса и соответствующего факта в генеративную ML-модель, тем самым вызывая формирование генеративной ML-моделью текстового представления соответствующего используемого машинно-формируемого ответа в ответ на используемый человеческий запрос с учетом контекста соответствующего факта.

[0023] В некоторых реализациях способа каждая из ML-модели семантического

подобия и генеративной ML-модели представляет собой ML-модель на основе Трансформера (Tranformer-based модель).

[0024] В некоторых реализациях способа генеративная ML-модель лишена части кодера ML-модели на основе Трансформера.

[0025] В некоторых реализациях способа система чат-бота является одной из (i) системы чат-бота для преобразования текста в текст; (ii) системы чат-бота для преобразования текста в речь; (iii) система чат-бота для преобразования речи в текст; и (iv) система чат-бота для преобразования речи в речь.

[0026] В соответствии со вторым широким аспектом настоящей технологии предоставляется сервер для обучения системы чат-бота формировать машинно-формируемые ответы на запросы пользователей системы чат-бота. Система чат-бота включает в себя: (i) модель машинного обучения (ML) семантического подобия, выполненную возможностью идентификации соответствующих фактов, имеющих отношение к запросам пользователей; и (ii) генеративную ML-модель, подлежащую обучению для формирования текстовых представлений машинно-формируемых ответов на запросы пользователей на основе запросов пользователей и соответствующих фактов. Сервер содержит: (i) процессор и (ii) не-кратковременный считываемый компьютером носитель, хранящий инструкции. Процессор при выполнении инструкций выполнен с возможностью: сбора диалоговых данных, включающих в себя (i) текстовые представления человеческих запросов диалогов пользователей на естественном языке; и (ii) текстовые представления соответствующих человеческих ответов диалогов, отвечающих на человеческие запросы; получения данные о фактах, включающие в себя текстовые представления фактов; идентификации из данных о фактах, используя ML-модель семантического подобия, для заданной диалоговой пары, включающей в себя текстовые представления заданного человеческого запроса и соответствующий человеческий ответ, отвечающий на него, из диалоговых данных, текстового представления соответствующего факта, относящегося к по меньшей мере одному из заданного человеческого запроса и соответствующего человеческого ответа; формирования обучающего набора данных, включающего в себя множество обучающих цифровых объектов, каждый из которых включает в себя: (i) текстовое представление заданного человеческого запроса; (ii) текстовое представление соответствующего факта; и (iii) соответствующую метку, являющуюся текстовым представлением соответствующего человеческого ответа на заданный человеческий запрос; подачи обучающего набора данных в систему чат-бота посредством: для заданного обучающего цифрового объекта из множества обучающих цифровых объектов, подачи процессором объединения текстового представления заданного человеческого запроса и текстового представления соответствующего факта в генеративную ML-модель для формирования текстового представления соответствующего машинно-формируемого ответа на заданный человеческий запрос с учетом контекста соответствующего факта; и; и оптимизации разницы между соответствующим машинно-формируемым ответом и соответствующим человеческим ответом на заданный человеческий запрос, за счет чего происходит обучение системы чат-бота формировать машинно-формируемые ответы на запросы пользователей.

[0027] В некоторых реализациях сервера для идентификации соответствующего факта для заданной диалоговой пары текстовых представлений с использованием ML-модели семантического подобия процессор выполнен с возможностью: подачи каждого из: (i) текстовых представлений человеческих запросов; (ii) текстовых представлений человеческих ответов; и (iii) текстовых представлений фактов в ML-модель

семантического подобия для формирования соответствующих векторных вложений каждого из (i) текстовых представлений человеческих запросов; (ii) текстовых представлений человеческих ответов; и (iii) текстовых представлений фактов; отображать соответствующие векторные вложения в векторное пространство; и идентифицируют соответствующий факт, относящийся к по меньшей мере одному из заданного человеческого запроса и соответствующего человеческого ответа, как факт, связанный с соответствующим векторным вложением, которое наиболее близко к соответствующему векторному вложению по меньшей мере одного из заданного человеческого запроса и соответствующего человеческого ответа в векторном пространстве.

[0028] В некоторых реализациях сервера для идентификации соответствующего факта процессор дополнительно выполнен с возможностью применения алгоритма k-ближайших соседей.

[0029] В некоторых реализациях сервера для идентификации соответствующего факта процессор дополнительно выполнен с возможностью применения функции ранжирования.

[0030] В некоторых реализациях сервера для оптимизации разницы процессор выполнен с возможностью оптимизации функции потерь, представляющей разницу между соответствующим машинно-формируемым ответом и соответствующим человеческим ответом.

[0031] В некоторых реализациях сервера процессор дополнительно выполнен с возможностью использования системы чат-бота для формирования машинно-формируемых ответов на запросы пользователей посредством: получения текстового представления используемого человеческого запроса заданного пользователя; идентификации из данных о фактах с использованием ML-модели семантического подобия соответствующего факта, относящегося к используемому человеческому запросу; и подачи процессором объединения текстовых представлений используемого человеческого запроса и соответствующего факта в генеративную ML-модель, тем самым вызывая формирование генеративной ML-моделью текстового представления соответствующего используемого машинно-формируемого ответа в ответ на используемый человеческий запрос с учетом контекста соответствующего факта.

[0032] В некоторых реализациях сервера каждая из ML-модели семантического подобия и генеративной ML-модели представляет собой ML-модель на основе Трансформера.

[0033] В некоторых реализациях сервера генеративная ML-модель лишена части кодера ML-модели на основе Трансформера.

[0034] В некоторых реализациях сервера система чат-бота является одной из (i) системы чат-бота для преобразования текста в текст; (ii) система чат-бота для преобразования текста в речь; (iii) система чат-бота для преобразования речи в текст; и (iv) система чат-бота для преобразования речи в речь.

[0035] В контексте настоящего описания модель «трансформера» (англ. transformer) представляет собой модель, имеющую архитектуру кодер-декодер, которая использует механизмы внимания. Механизмы внимания могут использоваться во время обработки данных кодером, во время обработки данных декодером и во время взаимодействия кодер-декодер. Различные механизмы внимания могут использоваться как часть модели трансформера.

[0036] Самостоятельное внимание может быть одним из компонентов модели трансформера. Разница между механизмом внимания и механизмом самовнимания

заключается в том, что самовнимание работает между представлениями одной природы: например, всеми состояниями кодера на каком-то слое. Механизм самовнимания является частью модели трансформера, где токены взаимодействуют друг с другом. Каждый токен в некотором смысле «смотрит» на другие токены в предложении с помощью механизма внимания, собирает контекст и обновляет предыдущее представление «самого себя». Каждый входной токен в механизме самовнимания получает три представления: (i) запрос, (ii) ключ и (iii) значение. Запрос используется, когда токен смотрит на других - он ищет информацию, чтобы лучше понять себя. Ключ отвечает на запрос запроса: он используется для вычисления веса внимания. Значение используется для вычисления вывода внимания: оно дает информацию о токенах, которые «говорят», что им это нужно (т.е. присваивают этому токену большие веса).

[0037] Маскированное самовнимание может быть еще одним компонентом модели трансформера. Декодер обычно включает в себя этот конкретный механизм самовнимания, который отличается от механизма самовнимания в кодере. В то время как кодер получает все токены сразу и токены могут просматривать все токены во входном предложении, в декодере токены формируются по одному - при формирования модель не знает, какие токены будут сформированы в дальнейшем. Чтобы запретить декодеру «заглядывать вперед», модель трансформера использует маскированное самовнимание, т. е. будущие токены маскируются.

[0038] Многоголовое внимание - еще одна из составляющих модели трансформера. Следует отметить, что понимание роли слова в предложении требует понимания того, как оно связано с разными частями предложения. Это важно не только при обработке исходного предложения, но и при создании целей. В результате этот тип механизма внимания может позволить модели трансформера «сосредоточиться на разных вещах». Вместо одного механизма внимания многоголовое внимание имеет несколько «голов», которые работают независимо. Это может быть реализовано в виде нескольких механизмов внимания, результаты которых объединяются.

[0039] Кодер модели трансформера может включать в себя механизм самовнимания кодера и блок сети прямого распространения (feedforward, FF). Механизм самовнимания кодера может быть механизмом многоголового внимания, используемым для того, чтобы токены «смотрели» друг на друга. Запросы, ключи, значения вычисляются из состояний кодера. Блок сети прямого распространения получает информацию от токенов и обрабатывает эту информацию.

[0040] Декодер модели трансформера может включать в себя механизм самовнимания декодера (маскированный), механизм внимания декодера-кодера и сеть прямого распространения. Механизм маскированного самовнимания декодера может быть маскированным механизмом многоголового внимания, используемым для токенов, чтобы «просматривать» предыдущие токены. Запросы, ключи, значения вычисляются из состояний декодера. Механизм внимания декодера-кодера может быть механизмом многоголового внимания, используемым для целевых токенов для «просмотра» исходной информации. Запросы вычисляются из состояний декодера, а ключи и значения вычисляются из состояний кодера. Блок сети прямого распространения получает информацию от токенов и обрабатывает эту информацию.

[0041] Можно сказать, что в кодере токены взаимодействуют друг с другом и обновляют свои представления. Также можно сказать, что в декодере целевой токен сначала просматривает ранее сформированные целевые токены, затем источник и, наконец, обновляет свое представление. Это можно повторить в несколько слоев. В одном неограничивающем осуществлении это может быть повторено 6 раз.

[0042] Как упоминалось выше, в дополнение к механизму внимания, данный слой имеет блок сети прямого распространения. Например, блок сети прямого распространения может быть представлен двумя линейными слоями с нелинейностью ReLU между ними. После просмотра других токенов с помощью механизма внимания модель использует блок сети прямого распространения для обработки этой новой информации. Модель трансформера может дополнительно содержать остаточные соединения для добавления входа блока к его выходу. Остаточные соединения можно использовать для стекирования слоев. В модели трансформера остаточные соединения могут использоваться после соответствующего механизма внимания и блока сети прямого распространения. Например, слой «Добавить и нормализовать» (англ. Add & Norm) может быть снабжен (i) вводом механизма внимания через остаточное соединение и (ii) выходом механизма внимания. Результат этого слоя «Добавить и нормализовать затем может быть предоставлен блоку сети прямого распространения» или другому механизму внимания. В другом примере слой «Добавить и нормализовать» может быть снабжен (i) входом блока сети прямого распространения через остаточное соединение и (ii) выходом блока сети прямого распространения. Как упоминалось выше, модель трансформера может содержать слои «Добавить и нормализовать». Говоря в общем, такой слой может независимо нормализовать векторное представление каждого примера в пакете - это делается для управления «перетеканием» к следующему слою. Нормализация слоя может улучшить стабильность сходимости, а иногда даже качество.

[0043] В контексте настоящего описания «сервер» представляет собой компьютерную программу, которая работает на надлежащем аппаратном обеспечении и способна принимать запросы (например, от клиентских устройств) по сети и выполнять эти запросы, или вызывать выполнение этих запросов. Аппаратное обеспечение может быть одним физическим компьютером или одной физической компьютерной системой, но ни то, ни другое не является обязательным для настоящей технологии. В настоящем контексте использование выражения "сервер" не предполагает, что каждая задача (например, принятые инструкции или запросы) или какая-либо конкретная задача будут приняты, выполнены или вызваны для выполнения одним и тем же сервером (т.е. тем же самым программным обеспечением и/или аппаратным обеспечением); данное выражение предполагает, что любое количество программных элементов или аппаратных устройств может быть задействовано в приеме/отправке, выполнении или вызове для выполнения любой задачи или запроса, или последствий любой задачи или запроса; и все это программное обеспечение и аппаратное обеспечение может быть одним сервером или многочисленными серверами, при этом оба данных случая включены в выражение "по меньшей мере один сервер".

[0044] В контексте настоящего описания "клиентское устройство" представляет собой любое компьютерное оборудование, которое способно выполнять программное обеспечение, которое является надлежащим для релевантной поставленной задачи. Таким образом, некоторые (неограничивающие) примеры клиентских устройств включают в себя персональные компьютеры (настольные компьютером, ноутбуки, нетбуки и т.д.), смартфоны и планшеты, а также сетевое оборудование, такое как маршрутизаторы, коммутаторы и шлюзы. Следует отметить, что устройство, выступающее в качестве клиентского устройства в настоящем контексте, не исключается из возможности выступать в качестве сервера для других клиентских устройств. Использование выражения "клиентское устройство" не исключает использования многочисленных клиентских устройств при приеме/отправке, выполнении или вызове для выполнения какой-либо задачи или запроса, или последствий любой задачи или

запроса, или этапов любого описанного в данном документе способа.

[0045] В контексте настоящего описания "база данных" представляет собой любую структурированную совокупность данных, независимо от ее конкретной структуры, программное обеспечение для администрирования базы данных, или компьютерное оборудование, на котором данные хранятся, реализуются или их делают доступными для использования иным образом. База данных может находиться на том же оборудовании, что и процесс, который хранит или использует информацию, хранящуюся в базе данных, или она может находиться на отдельном оборудовании, например на выделенном сервере или множестве серверов.

[0046] В контексте настоящего описания выражение "информация" включает в себя информацию любого характера или вида, который способен храниться в базе данных любым образом. Эта информация включает в себя, помимо прочего, аудиовизуальные произведения (изображения, фильмы, звукозаписи, презентации и т. д.), данные (данные о местоположении, числовые данные и т. д.), текст (мнения, комментарии, вопросы, сообщения и т. д.), документы, электронные таблицы, списки слов и т. д.

[0047] В контексте настоящего описания, если специально не указано иное, подразумевается, что термин «компонент» включает в себя программное обеспечение (соответствующее конкретному аппаратному контексту), которое является как необходимым, так и достаточным для реализации конкретной функции (функций), на которую ссылаются.

[0048] В контексте настоящего описания предполагается, что выражение "используемый компьютером носитель хранения информации" включает в себя носители любого характера и вида, в том числе RAM, ROM, диски (CD-ROM, DVD, дискеты, накопители на жестких дисках и т.д.), USB-ключи, твердотельные накопители, ленточные накопители и т.д.

[0049] В контексте настоящего описания слова "первый", "второй", "третий" и т.д. используются в качестве прилагательных только для того, чтобы позволить отличать существительные, к которым они относятся, друг от друга, а не для описания какой-либо особой взаимосвязи между такими существительными. Таким образом, например, следует понимать, что использование терминов «первый сервер» и «третий сервер» не подразумевает какой-либо конкретный порядок, тип, хронологию, иерархию или ранжирование (например) сервера/между серверами, а также их использование (само по себе) не подразумевает, что какой-либо «второй сервер» обязательно должен существовать в любой данной ситуации. Кроме того, как обсуждается в других контекстах данного документа, ссылка на "первый" элемент и "второй" элемент не исключает того, что эти два элемента фактически являются одним и тем же элементом реального мира. Таким образом, например, в некоторых случаях "первый" сервер и "второй" сервер могут быть одним и тем же программным обеспечением и/или аппаратным обеспечением, в других случаях они могут представлять собой разное программное обеспечение и/или аппаратное обеспечение.

[0050] Каждая из реализаций настоящей технологии обладает по меньшей мере одним из вышеупомянутых аспектов и/или цели, но не обязательно имеет их все. Следует понимать, что некоторые аспекты настоящей технологии, которые возникли в попытке достичь вышеупомянутой цели, могут не удовлетворять этой цели и/или удовлетворять другим целям, которые не описаны в данном документе явным образом.

[0051] Дополнительные и/или альтернативные признаки, аспекты и преимущества реализаций настоящей технологии станут очевидными из нижеследующего описания, сопроводительных чертежей и приложенной формулы изобретения.

КРАТКОЕ ОПИСАНИЕ ЧЕРТЕЖЕЙ

[0052] Для лучшего понимания настоящей технологии, а также других аспектов и ее дополнительных признаков, ссылка приводится на нижеследующее описание, которое должно использоваться в сочетании с сопроводительными чертежами, на которых:

[0053] На фиг. 1 изображена принципиальная схема примерной компьютерной системы для реализации некоторых неограничивающих вариантов осуществления систем и/или способов настоящей технологии;

[0054] На фиг. 2 изображена сетевая вычислительная среда, подходящая для некоторых реализаций определенных неограничивающих вариантов осуществления настоящей технологии;

[0055] На фиг. 3 изображена схематическая диаграмма процесса формирования системой чат-бота, размещенной на сервере, присутствующем в сетевой вычислительной среде, показанной на Фиг. 2, текстового представления соответствующего машинно-формируемого ответа на заданный человеческий запрос в соответствии с неограничивающими вариантами осуществления настоящей технологии;

[0056] На фиг. 4 изображена схематическая диаграмма архитектуры модели машинного обучения, пригодной для использования в некоторых неограничивающих реализациях настоящей технологии;

[0057] На фиг. 5 изображена схематическая диаграмма предварительного этапа обучения сервером, присутствующим в сетевой вычислительной среде, показанной на Фиг. 2, генеративной модели системы чат-бота для формирования текстового представления соответствующего машинно-формируемого ответа на заданный человеческий запрос в соответствии с некоторыми неограничивающими вариантами осуществления настоящей технологии;

[0058] На Фиг. 6 изображена схематическая диаграмма этапа точной настройки сервера, присутствующего в сетевой вычислительной среде, изображенной на Фиг. 2 генеративной модели системы чат-бота для формирования текстового представления соответствующего машинно-формируемого ответа в соответствии с некоторыми неограничивающими вариантами осуществления настоящей технологии;

[0059] На фиг. 7 изображено схематическое представление использования сервером, присутствующим в сетевой вычислительной среде, показанной на Фиг. 2, генеративной модели системы чат-бота для формирования текстового представления соответствующего машинно-формируемого ответа в соответствии с некоторыми неограничивающими вариантами осуществления настоящей технологии; и

[0060] На фиг. 8 изображена блок-схема способа обучения с помощью сервера, присутствующего в сетевой вычислительной среде, показанной на фиг. 2, и системы чат-бота, размещенной таким образом, для формирования текстового представления соответствующего машинно-формируемого ответа в соответствии с неограничивающими вариантами осуществления настоящей технологии.

ПОДРОБНОЕ ОПИСАНИЕ

[0061] Приведенные в данном документе примеры и условные формулировки призваны главным образом помочь читателю понять принципы настоящей технологии, а не ограничить ее объем такими конкретно приведенными примерами и условиями. Должно быть понятно, что специалисты в данной области техники смогут разработать различные механизмы, которые, хоть и не описаны в данном документе явным образом, тем не менее воплощают принципы настоящей технологии и включаются в ее суть и объем.

[0062] Кроме того, нижеследующее описание может описывать реализации настоящей

технологии в относительно упрощенном виде для целей упрощения понимания. Специалисты в данной области техники поймут, что различные реализации настоящей технологии могут иметь большую сложность.

[0063] В некоторых случаях также могут быть изложены примеры модификаций настоящей технологии, которые считаются полезными. Это делается лишь для содействия пониманию и, опять же, не для строгого определения объема или очерчивания границ настоящей технологии. Эти модификации не являются исчерпывающим списком, и специалист в данной области техники может осуществлять другие модификации, все еще оставаясь при этом в рамках объема настоящей технологии. Кроме того, случаи, когда примеры модификаций не приводятся, не следует толковать так, что никакие модификации не могут быть осуществлены и/или что описанное является единственным способом реализации такого элемента настоящей технологии.

[0064] Кроме того, все содержащиеся в данном документе утверждения, в которых указываются принципы, аспекты и реализации настоящей технологии, а также их конкретные примеры, призваны охватить как структурные, так и функциональные эквиваленты, вне зависимости от того, известны ли они в настоящее время или будут разработаны в будущем. Таким образом, например, специалисты в данной области техники осознают, что любые блок-схемы в данном документе представляют концептуальные виды иллюстративной схемы, воплощающей принципы настоящей технологии. Аналогичным образом, будет понятно, что любые блок-схемы, схемы последовательности операций, схемы изменения состояний, псевдо-коды и подобное представляют различные процессы, которые могут быть по сути представлены на считываемых компьютерам носителях и исполнены компьютером или процессором вне зависимости от того, показан такой компьютер или процессор явным образом или нет.

[0065] Функции различных элементов, показанных на фигурах, в том числе любого функционального блока, помеченного как «процессор» или «графический процессор», могут быть обеспечены с помощью специального аппаратного обеспечения, а также аппаратного обеспечения, способного исполнять программное обеспечение, в сочетании с надлежащим программным обеспечением. При обеспечении процессором функции могут быть обеспечены одним выделенным процессором, одним совместно используемым процессором и/или множеством отдельных процессоров, некоторые из которых могут быть совместно используемыми. В некоторых вариантах осуществления настоящей технологии процессор может быть процессором общего назначения, таким как центральный процессор (CPU), или процессором, выделенным для конкретной цели, таким как графический процессор (GPU). Кроме того, явное использование термина «процессор» или «контроллер» не должно истолковываться как относящееся исключительно к аппаратному обеспечению, способному исполнять программное обеспечение, и может в неявной форме включать в себя, без ограничений перечисленным, аппаратное обеспечение цифрового сигнального процессора (DSP), сетевой процессор, интегральную схему специального назначения (ASIC), программируемую пользователем вентильную матрицу (FPGA), постоянную память (ROM) для хранения программного обеспечения, оперативную память (RAM) и/или энергонезависимый (не-кратковременный) носитель данных. Другое аппаратное обеспечение, традиционное и/или специализированное, также может быть включено в состав.

[0066] Программные модули, или просто модули, в качестве которых может подразумеваться программное обеспечение, могут быть представлены в настоящем документе как любая комбинация элементов блок-схемы последовательности операций

или других элементов, указывающих выполнение этапов процесса и/или текстовое описание. Такие модули могут выполняться аппаратным обеспечением, которое явно или неявно показано.

[0067] Учитывая эти основополагающие вещи, рассмотрим некоторые неограничивающие примеры, чтобы проиллюстрировать различные реализации аспектов настоящей технологии.

Компьютерная система

[0068] Со ссылкой на Фиг. 1, здесь изображена компьютерная система 100, подходящая для использования с некоторыми реализациями настоящей технологии. Компьютерная система 100 содержит различные аппаратные компоненты, в том числе один или более одноядерных или многоядерных процессоров, совместно представленных процессором 110, графическим процессором (GPU) 111, твердотельным накопителем 120, оперативной памятью 130, интерфейсом 140 дисплея и интерфейсом 150 ввода/вывода.

[0069] Связь между различными компонентами компьютерной системы 100 может быть обеспечена одной или более внутренними и/или внешними шинами 160 (например, шиной PCI, универсальной последовательной шиной, шиной Firewire IEEE 1394, шиной SCSI, шиной Serial-ATA и т.д.), с которой различные компоненты оборудования связаны электронным образом.

[0070] Интерфейс 150 ввода/вывода может быть связан с сенсорным экраном 190 и/или с одной или более внутренними и/или внешними шинами 160. Сенсорный экран 190 может быть частью дисплея. В некоторых вариантах осуществления, сенсорный экран 190 является дисплеем. В вариантах осуществления, проиллюстрированных на фиг. 1, сенсорный экран 190 содержит сенсорное аппаратное обеспечение 194 (например, чувствительные к давлению ячейки, встроенные в слой дисплея, позволяющие обнаруживать физическое взаимодействие между пользователем и дисплеем) и контроллер 192 сенсорного ввода/вывода, обеспечивающий возможность связи с интерфейсом 140 дисплея и/или одной или более внутренними и/или внешними шинами 160. В некоторых вариантах осуществления интерфейс 150 ввода/вывода может быть подключен к клавиатуре (не показана), мыши (не показана) или трекпаду (не показан), позволяющим пользователю взаимодействовать с компьютерной системой 100 в дополнение к или вместо сенсорного экрана 190. В некоторых вариантах осуществления компьютерная система 100 может содержать один или более микрофонов (не показаны). Микрофоны могут записывать аудио, например, пользовательские высказывания. Пользовательские высказывания могут быть преобразованы в команды для управления компьютерной системой 100.

[0071] Следует отметить, что некоторые компоненты компьютерной системы 100 могут быть опущены в некоторых неограничивающих вариантах осуществления настоящей технологии. Например, сенсорный экран 190 может отсутствовать, особенно (но не ограничиваясь этим), когда компьютерная система реализована как интеллектуальное устройство динамика.

[0072] Согласно реализациям настоящей технологии, твердотельный накопитель 120 хранит программные инструкции, подходящие для загрузки в оперативную память 130 и исполнения процессором 110 и/или GPU 111. Например, программные инструкции могут быть частью библиотеки или приложения.

Сетевая вычислительная среда

[0073] Со ссылкой на фиг. 2, здесь изображена схема сетевой вычислительной среды 200, подходящая для использования с некоторыми вариантами осуществления систем

и/или способов настоящей технологии. Сетевая вычислительная среда 200 содержит сервер 202, связанный с возможностью передачи данных через сеть 208 связи с электронным устройством 204. В неограничивающих вариантах осуществления настоящей технологии электронное устройство 204 может быть связано с пользователем 206.

[0074] В некоторых неограничивающих вариантах осуществления настоящей технологии сервер 202 реализован как обычный компьютерный сервер и может содержать некоторые или все компоненты компьютерной системы 100 по фиг. 1. В одном неограничивающем примере сервер 202 реализуется как сервер Dell™ PowerEdge™, работающий под управлением операционной системы Microsoft™ Windows Server™, но также может быть реализован в любом другом подходящем аппаратном обеспечении, программном обеспечении и/или микропрограммном обеспечении, или их комбинации. В проиллюстрированном неограничивающем варианте осуществления настоящей технологии сервер 202 является единственным сервером. В альтернативных неограничивающих вариантах осуществления настоящей технологии (не показаны) функциональные возможности сервера 202 могут быть распределены и могут быть реализованы через несколько серверов.

[0075] Кроме того, электронное устройство 204 может представлять собой любое компьютерное оборудование, способное запускать программное обеспечение, подходящее для соответствующей текущей задачи. Таким образом, некоторые неограничивающие примеры электронного устройства 204 могут включать в себя персональные компьютеры (настольные компьютеры, ноутбуки, нетбуки и т. д.), смартфоны и планшеты. С этой целью в некоторых неограничивающих вариантах осуществления настоящей технологии электронное устройство 204 может также содержать некоторые или все компоненты компьютерной системы 100, изображенной на фиг. 1.

[0076] В соответствии с некоторыми неограничивающими вариантами осуществления настоящей технологии сетевая вычислительная среда 200 может быть выполнена с возможностью обеспечения и/или поддержания автоматической связи с пользователем 206, как будет описано ниже. С этой целью, в соответствии с некоторыми неограничивающими вариантами осуществления настоящей технологии, сервер 202 может быть выполнен с возможностью выполнения системы 210 чат-бота (также называемой здесь «приложением виртуального помощника»).

[0077] Говоря в общем, система 210 чат-бота выполнена с возможностью инициирования и дальнейшего поддержания автоматической связи с пользователем 206, помогая ему получать различные услуги, предоставляемые соответствующими поставщиками услуг. Например, в некоторых неограничивающих вариантах осуществления настоящей технологии данная онлайн-служба может содержать платформу потоковой передачи видео (такую как платформа потоковой передачи видео Netflix™, платформа потоковой передачи видео Kinopoisk™ и т.п.), а система 210 чат-бота посредством автоматической связи с пользователем 206 может быть выполнена с возможностью помощи пользователю 206 в поиске контента платформы потоковой передачи видео, например, путем предоставления рекомендаций по контенту. В другом примере данная онлайн-служба может быть поисковой системой (такой как поисковая система Yandex™, поисковая система Google™ и т.п.); и система 210 чат-бота может быть выполнена с возможностью предоставления помощи пользователю 206 в веб-поиске, чтобы идентифицировать более релевантные результаты поиска для запросов, отправленных пользователем 206. В еще одном примере данная онлайн-служба может

быть платформой онлайн-листинга (например, платформой онлайн-листинга Yandex™ Market и платформой онлайн-листинга Avito™, платформой онлайн-листинга Kijiji™ и т.п.); и система 210 чат-бота может быть выполнена с возможностью помогать пользователю 206 в навигации по товарам, доступным для покупки на онлайн-платформе листинга, для определения интересующих товаров. Кроме того, в этом примере система 210 чат-бота может быть выполнена с возможностью реализации определенных функций обслуживания клиентов, таких как прием и управление заказами, обработка жалоб и т.п. В еще других неограничивающих вариантах осуществления настоящей технологии данная онлайн-служба может быть банковским интернет-приложением банка, а система 210 чат-бота может быть выполнена с возможностью помощи пользователю 206 в получении определенных банковских услуг, таких как проведение транзакций, получение кредитов, осуществление платежей и т.п.

[0078] В некоторых неограничивающих вариантах осуществления настоящей технологии объект, владеющий системой 210 чат-бота и данной онлайн-службой, может быть одним и тем же, и данная онлайн-служба может выполняться сервером 202. Однако в других неограничивающих вариантах осуществления настоящей технологии система 210 чат-бота и данная онлайн-служба могут принадлежать разным объектам; в связи с этим данная онлайн-служба может выполняться сторонним сервером (не показан), который имеет доступ к системе 210 чат-бота, управляемой сервером 202, через сеть 208 связи.

[0079] Однако в некоторых неограничивающих вариантах осуществления настоящей технологии система 210 чат-бота может быть выполнена с возможностью помощи пользователю 206 в получении автономных услуг. Например, данная автономная услуга может быть рестораном; и система 210 чат-бота может быть выполнена с возможностью помощи пользователю 206 в бронировании столиков или заказе еды для доставки. В другом примере данная автономная служба может быть банком, а система 210 чат-бота может быть выполнена с возможностью помощи пользователю 206 в получении определенных банковских услуг, таких как проведение транзакций, получение кредитов, осуществление платежей и т.п.

[0080] В еще других неограничивающих вариантах осуществления настоящей технологии система 210 чат-бота может быть автономной системой чат-бота, работающей на электронном устройстве 204. В этих вариантах осуществления система 210 чат-бота может быть выполнена с возможностью работы независимо от любой из вышеупомянутых служб; однако может помочь пользователю 206 перемещаться по различным приложениям, предварительно установленным на электронном устройстве 204, которые могут быть связаны с некоторыми из упомянутых выше услуг.

[0081] В конкретном неограничивающем примере система 210 чат-бота может быть реализована как приложение виртуального помощника ALISA™, предоставленное ООО «ЯНДЕКС» по адресу: 119021, Москва, ул. Льва Толстого, д. 16, Россия. Однако следует отметить, что система 210 чат-бота может быть реализована как любое другое коммерческое или проприетарное приложение виртуального помощника.

[0082] В тех вариантах осуществления, где система 210 чат-бота предварительно установлена на электронном устройстве 204, пользователь 206 может активировать систему 210 чат-бота, например, с помощью слова пробуждения, связанного с системой 210 чат-бота, такого как, например, «Привет, Siri» или «Привет Алиса». В других неограничивающих вариантах осуществления настоящей технологии пользователь 206 может активировать систему 210 чат-бота через приложение веб-браузера электронного устройства 204, отправив туда URL-адрес, связанный с браузерной версией системы

210 чат-бота. Кроме того, система 210 чат-бота может быть выполнена с возможностью: (i) приема индикации заданного человеческого запроса 212 от пользователя 206; и в ответ на это (ii) формирования соответствующий машинно-формируемый ответ 214, тем самым поддерживая связь с пользователем 206.

[0083] Не ограничивается то, как заданный человеческий запрос 212 может быть отправлен в систему 210 чат-бота и как система 210 чат-бота может быть дополнительно выполнена с возможностью формирования соответствующего машинно-формируемого ответа 214; и обычно зависит от конкретной реализации системы 210 чат-бота. Например, в соответствии с некоторыми неограничивающими вариантами осуществления настоящей технологии система 210 чат-бота может быть системой чат-бота преобразования текста в текст, которая выполнена с возможностью приема текстовых представлений (таких как текстовые строки) человеческих запросов и формирования текстовых представлений соответствующих машинно-формируемых ответов.

[0084] В других неограничивающих вариантах осуществления настоящей технологии система 210 чат-бота может быть реализована как голосовой системы система чат-бота, выполненная с возможностью приема человеческих запросов в виде человеческих высказываний и формирования соответствующих машинно-формируемых ответов в форме машинно-формируемых высказываний. В этих вариантах осуществления система 210 чат-бота может содержать: (i) модель преобразования речи в текст (STT) (не показана), выполненная с возможностью обработки разговорного человеческого языка, такого как человеческое высказывание, представляющее заданный человеческий запрос 212, и распознавания в нем отдельных слов, тем самым формируя текстовое представление (такое как текстовая строка) заданного человеческого запроса 212 для дальнейшей обработки; и (ii) модель преобразования текста в речь (TTS) (не показана), выполненная с возможностью преобразования машинно-формируемого текста, такого как текстовое представление соответствующего машинно-формируемого ответа 214, в экземпляры речи на естественном языке для дальнейшего воспроизведения говорящим на электронном устройстве 204. Излишне упоминать, что в этих вариантах осуществления, чтобы позволить системе 210 чат-бота принимать человеческие высказывания, электронное устройство 204 может включать в себя дополнительные компоненты, такие как микрофон (не показан отдельно) для преобразования принятых звуков, записанных вблизи электронного устройства 204, в считываемый компьютером формат, такой как цифровой аудиоформат, включая, например, MP3, Ogg и т.п.; и динамик (также отдельно не изображенный) для воспроизведения входящих аудиосигналов вблизи электронного устройства 204. Кроме того, для воспроизведения машинно-формируемых высказываний, сформированных моделью TTS, электронное устройство 204 может включать в себя динамик.

[0085] Способы реализации моделей STT и TTS не ограничены. В некоторых неограничивающих вариантах осуществления настоящей технологии каждая из моделей STT и TTS может быть реализована как нейронные сети (NN) на основе Трансформера, такие как архитектура 400 модели машинного обучения, описанная ниже со ссылкой на фиг. 4. В этих вариантах осуществления модели STT и TTS могут быть обучены для дальнейшего использования с системой 210 чат-бота, как описано, например, в заявке на патент США, находящейся в совместном владении: 18/081,634, поданной 14 декабря 2022 г. и озаглавленной ***«СПОСОБ И СИСТЕМА РАСПОЗНАВАНИЯ ПОЛЬЗОВАТЕЛЬСКОГО ВЫСКАЗЫВАНИЯ»***, содержание которой полностью включено в настоящий документ посредством ссылки.

[0086] Гибридные реализации системы 210 чат-бота, такие как преобразование текста

в голос и голос в текст, также предусмотрены, не выходя за рамки настоящей технологии.

[0087] Таким образом, в некоторых неограничивающих вариантах осуществления настоящей технологии система 210 чат-бота может быть выполнена с возможностью приема текстового представления заданного человеческого запроса 212, например, *«Пожалуйста, порекомендуйте хороший научно-фантастический фильм»*. Кроме того, система 210 чат-бота может быть выполнена с возможностью формирования текстового представления соответствующего машинно-формируемого ответа 214, например: *«Это то, что я нашел по вашему запросу»*, наряду с этим система 210 чат-бота может быть выполнена с возможностью извлечения списка рекомендуемых научно-фантастических фильмов, идентифицированных, например, поисковой системой или платформой потокового видео. Кроме того, сервер 202 может быть выполнен с возможностью передачи текстового представления соответствующего машинно-формируемого ответа 214 вместе со списком рекомендуемых научно-фантастических фильмов на электронное устройство 204 для представления пользователю 206.

[0088] Со ссылкой на фиг. 3, здесь изображена схема процесса формирования сервером 202, на котором запущена система 210 чат-бота, соответствующего машинно-формируемого ответа 214 на заданный человеческий запрос 212 в соответствии с некоторыми неограничивающими вариантами осуществления настоящей технологии.

[0089] В соответствии с некоторыми неограничивающими вариантами осуществления настоящей технологии система 210 чат-бота может содержать модель 302 вложения, выполненную с возможностью формирования векторных вложений текстовых представлений пользовательских запросов, таких как соответствующее векторное вложение 306 заданного человеческого запроса 212. Подробное описание модели 302 вложения будет представлено ниже; однако, говоря в общем, модель 302 вложения выполнена с возможностью разложения текстовых входных данных, таких как, например, слова, предложения или абзацы, в числовые векторы заданного пространства векторных вложений, так что векторные вложения семантически подобных текстовых входных данных, то есть более близких по смыслу, расположены в заданном пространстве векторных вложений ближе друг к другу, чем те, которые семантически различны, то есть имеют разные значения. С этой целью модель 302 вложения также упоминается здесь как «модель семантического подобия».

[0090] Кроме того, в соответствии с некоторыми неограничивающими вариантами осуществления настоящей технологии система 210 чат-бота может содержать генеративную модель 304, выполненную с возможностью формирования на основе соответствующего векторного вложения 306 заданного человеческого запроса 212 текстового представления соответствующего машинно-формируемого ответа 214.

[0091] Говоря в общем, генеративная модель 304 представляет собой модель машинного обучения (такую как модель машинного обучения на основе глубокой нейронной сети), которая была обучена формировать текстовые представления машинно-формируемых ответов на запросы пользователей на основе обучающего набора данных, включающего в себя текстовые представления множества обучающих диалоговых пар, включающих в себя (i) обучающие запросы пользователей; и (ii) соответствующие обучающие ответы на них на естественном языке, таком как, например, русский или английский.

[0092] Например, сервер 202 может быть выполнен с возможностью получения таких диалоговых пар «запросы-ответы» для обучающего набора данных путем обхода различных ресурсов сети 208 связи, предоставляющих доступ к общедоступным

разговорам людей, таким как комментарии или сообщения на данном веб-сайте социальной сети или обсуждения на данном веб-сайте-форуме. В другом примере сервер 202 может быть выполнен с возможностью получения этих диалоговых пар на естественном языке от автономных поставщиков услуг, которые могут включать в себя диалоговые строки между сотрудниками, такими как представители отдела обслуживания клиентов, поставщиков услуг и их клиентов.

[0093] Однако, как упоминалось выше, генеративные модели машинного обучения, обученные исключительно на диалоговых парах «запросы-ответы», имеют тенденцию давать машинно-формируемые ответы, которые могут быть восприняты пользователями системы 210 чат-бота, такими как пользователь 206, как неинтересные (или иным образом «плоские») или не относящиеся к моменту времени, когда пользователь 206 отправляет заданный человеческий запрос 212 в систему 210 чат-бота, что может повлиять на общий пользовательский опыт пользователя 206 с системой 210 чат-бота.

[0094] В связи с этим разработчики настоящей технологии осознали, что добавление фактического контекста к каждой из диалоговых пар может позволить обучить генеративную модель 304 формировать более естественные и разнообразные машинно-формируемые ответы на запросы пользователей. Более конкретно, настоящие способы и системы направлены на обучение системы 210 чат-бота формировать текстовые представления соответствующих дополненных машинно-формируемых ответов на запросы пользователей, таких как соответствующий дополненный машинно-формируемый ответ 714 на заданный человеческий запрос 212 (как будет подробно описано со ссылкой на фиг. 7), с учетом контекста соответствующих фактов, которые были идентифицированы как относящиеся к запросам пользователей.

[0095] Примерная архитектура, а также то, как сервер 202 может быть выполнен с возможностью обучения как модели 302 вложения, так и генеративной модели 304 для формирования текстового представления соответствующего дополненного машинно-формируемого ответа 714 на заданный человеческий запрос 212, в соответствии с некоторыми неограничивающими вариантами осуществления настоящей технологии, будут подробно описаны ниже со ссылкой на фиг. 4-7.

Сеть связи

[0096] В некоторых неограничивающих вариантах осуществления настоящей технологии сеть 208 связи представляет собой Интернет. В альтернативных неограничивающих вариантах осуществления настоящей технологии сеть 208 связи может быть реализована как любая подходящая локальная сеть (LAN), глобальная сеть (WAN), частная сеть связи или подобная. Следует четко понимать, что реализации для сети 208 связи предназначены только для иллюстративных целей. То, как реализуется соответствующий канал связи (не пронумерованный отдельно) между сервером 202, электронным устройством 204 и сетью 208 связи, будет зависеть, среди прочего, от того, как реализован каждый из сервера 202 и электронного устройства 204. Просто в качестве примера, а не ограничения, в тех вариантах осуществления настоящей технологии, где электронное устройство 204 реализовано как устройство беспроводной связи, такое как смартфон, линия связи может быть реализована как линия беспроводной связи. Примеры линий беспроводной связи включают в себя, но без ограничения, линию сети связи 3G, линию сети связи 4G и подобные. Сеть 208 связи может также использовать беспроводное соединение с сервером 202.

Архитектура модели машинного обучения

[0097] Со ссылкой на фиг. 4, здесь изображена архитектура 400 модели машинного обучения, пригодная для использования по меньшей мере с некоторыми

неограничивающими вариантами осуществления настоящей технологии. Архитектура 400 модели машинного обучения основана на архитектуре модели нейронной сети «Трансформер» (Transformer), как описано, например, в статье за авторством Vaswani и др., названная «*Attention Is All You Need*» и опубликованная в материалах 31-й конференции по системам обработки нейронной информации (NIPS 2017), содержание которой полностью включено в настоящий документ посредством ссылки.

[0098] Таким образом, архитектура 400 модели машинного обучения может содержать стек слоев 402 кодера и стек слоев 403 декодера, которые могут быть выполнены с возможностью обработки входных данных 412 и целевых данных 417 архитектуры 400 модели машинного обучения соответственно.

[0099] Кроме того, данный блок 404 кодера из стека слоев 402 кодера включает в себя слой 406 кодера с многоголовым вниманием (multi-head attention, MHA) и слой 408 NN прямого распространения кодера. Слой 406 MHA кодера включает в себя зависимости между частями входных данных 412, предоставленных ему. Например, если входные данные 412 включают в себя текстовые данные, такие как текстовое предложение, слой 406 MHA кодера может включать в себя зависимости между словами предложения. В другом примере, когда входные данные 412 для стека слоев 402 кодера включают в себя звуковой сигнал, такой как сигнал, представляющий человеческое высказывание, слой 406 MHA кодера может включать в себя зависимости между определенными звуками и/или акустическими характеристиками человеческого высказывания. Такие зависимости могут использоваться слоем 406 MHA кодера для определения контекстуальной информации заданной части входных данных 412 для стека слоев 402 кодера (например, представляющего данное слово предложения) относительно другой части входных данных 412.

[00100] Кроме того, слой 408 NN прямого распространения кодера выполнен с возможностью преобразования данных, вводимых в него с слоя 406 MHA кодера, в формат, который может быть получен одним или более последующими слоями архитектуры 400 модели машинного обучения, такими как слой 409 MHA кодера-декодера, как будет описано ниже. Слой 408 NN прямого распространения кодера обычно не зависит от слоя 406 MHA кодера, и поэтому входные данные для слоя 408 NN прямого распространения кодера могут обрабатываться параллельно.

[00101] Кроме того, входные данные 412 для стека слоев 402 кодера могут быть представлены множеством входных векторов 414, сформированных алгоритмом 410 входного вложения. Говоря в общем, алгоритм 410 входного вложения выполнен с возможностью формирования векторных вложений фиксированной размерности входных данных 412 в соответствующем пространстве векторных вложений. Другими словами, если входные данные 412 содержат текстовые данные, такие как текстовое предложение, алгоритм 410 входного вложения может быть выполнен с возможностью формирования множества входных векторов 414, где координаты векторных вложений, представляющих слова текстового предложения, схожие по смыслу, расположены ближе друг к другу в соответствующем пространстве вложений. Таким образом, алгоритм 410 входного вложения может быть реализован как алгоритм вложения текста, включая, помимо прочего, один из алгоритмов вложения текста Word2Vec, алгоритм вложения текста GloVe и т.п.

[00102] Таким образом, заданный один из множества входных векторов 414 может включать в себя числовые значения, например, 768 значений с плавающей запятой, представляющих соответствующую часть входных данных 412, таких как слово, часть заданного человеческого запроса 212 и т.п.

[00103] Кроме того, формирование множества входных векторов 414 может дополнительно включать в себя применение алгоритма позиционного вложения (не показан), выполненного с возможностью регистрации позиционной информации в частях входных данных 412. Например, если входные данные 412 включают в себя текстовое предложение, алгоритм позиционного вложения может быть выполнен с возможностью формирования вектора, указывающего позиционную информацию среди слов в этом текстовом предложении. Другими словами, алгоритм позиционного вложения может быть выполнен с возможностью формирования вектора, сохраняющего контекстную информацию во входных данных 412, которые могут быть дополнительно добавлены к множеству входных векторов 414. Способ реализации алгоритма позиционного вложения не ограничен; и может включать в себя, без ограничений перечисленным, алгоритм позиционного вложения синусоид, алгоритм позиционного вложения с наложением кадров и сверточный алгоритм позиционного вложения, в качестве примера.

[00104] Следует отметить, что стек слоев 402 кодера может включать в себя несколько блоков кодера, таких как, например, 6 или 12, реализованных аналогично данному блоку 404 кодера.

[00105] Кроме того, данный блок 405 декодера из стека слоев 403 декодера архитектуры 400 модели машинного обучения также включает в себя (i) слой 407 MHA декодера; и (ii) слой 411 NN прямого распространения декодера, который в целом может быть реализован аналогично слою 406 MHA кодера и слою 408 NN прямого распространения кодера соответственно. Однако архитектура заданного блока 405 декодера отличается от архитектуры заданного блока 404 кодера тем, что заданный блок 405 декодера дополнительно включает в себя слой 409 MHA кодера-декодера. Слой 409 MHA кодера-декодера выполнен с возможностью (i) приема входных векторов от стека слоев 402 кодера и от слоя 407 MHA декодера; и, таким образом, (ii) определения во время процесса обучения зависимости между входными данными 412 и целевыми данными 417 (например, текстовыми данными) архитектуры 400 модели машинного обучения, вводимой в стек слоев 403 декодера. Другими словами, выходные данные слоя 409 MHA кодера-декодера представляют собой векторы внимания, включающие в себя данные, указывающие на отношения между соответствующими частями входных данных 412 и целевыми данными 417.

[00106] Подобно входным данным 412, для подачи целевых данных 417 в заданный блок 405 декодера алгоритм 415 целевого вложения может быть применен к целевым данным 417 для формирования множества целевых векторов 419, содержащих числовые представления соответствующих частей целевых данных 417.

[00107] Как можно понять, в тех вариантах осуществления, где целевые данные 417 представляют собой текстовые данные, алгоритм 415 целевого вложения может быть реализован аналогично алгоритму 410 входного вложения. Кроме того, позиционный алгоритм также может быть применен к множеству целевых векторов 419 для регистрации позиционных данных среди частей целевых данных 417, как описано выше в отношении множества входных векторов 414.

[00108] Как станет очевидным из приведенного ниже описания, архитектура 400 модели машинного обучения может быть выполнена с возможностью приема входных данных 412 и целевых данных 417 от цифрового объекта, такого как один из заданного обучающего цифрового объекта 504 или заданный другой обучающий цифровой объект 604, как будет описано со ссылкой на фиг. 5 и 6 соответственно.

[00109] Аналогичным образом следует отметить, что стек слоев 403 декодера может

включать в себя несколько блоков декодера, таких как, например, 6 или 12, реализованных аналогично данному блоку 405 декодера. Кроме того, как можно понять, после обучения архитектуры 400 модели машинного обучения каждый блок стека слоев 402 кодера и стека слоев 403 декодера будет иметь разные веса, способствующие формированию выходных данных 425. Для корректировки весов в процессе обучения к архитектуре 400 модели машинного обучения может быть применен алгоритм обратного распространения ошибки, и может быть определена и дополнительно оптимизирована разница между входными данными 412 и выходными данными 425. Такая разница может быть выражена функцией потерь, такой как функция кросс-энтропийных потерь.

[00110] Следует четко понимать, что другие реализации функции потерь также предусмотрены неограничивающими вариантами осуществления настоящей технологии и могут включать в себя, в качестве примера, а не в качестве ограничения, функцию потерь среднеквадратичной ошибки (Mean Squared Error Loss function), функцию потерь Хьюбера (Huber Loss function), функцию потерь шарнира (Hinge Loss function) и другие.

[00111] Кроме того, не ограничивается то, каким образом сервер 202 может быть выполнен с возможностью оптимизации функции потерь, и в некоторых неограничивающих вариантах осуществления настоящей технологии это будет зависеть, как правило, от дифференцируемости функции потерь. Например, если функция потерь непрерывно дифференцируема, подходы к ее минимизации могут включать в себя, помимо прочего, алгоритм градиентного спуска, алгоритм оптимизации Ньютона и другие. В тех вариантах осуществления, где функция потерь является недифференцируемой, для ее минимизации сервер 202 может быть выполнен с возможностью применения, например, по меньшей мере одного из прямых алгоритмов, стохастических алгоритмов и алгоритмов популяции.

[00112] Выходные данные 425 архитектуры 400 модели машинного обучения могут включать в себя выходной вектор, соответствующий заданному одному из множества входных векторов 414 и/или множества целевых векторов 419. Например, как станет очевидным из приведенного ниже описания, в тех вариантах осуществления, где входные данные 412 для архитектуры 400 модели машинного обучения включают в себя текстовое представление заданного человеческого запроса 212, выходной вектор может включать в себя вероятности, указывающие на текстовое представление соответствующего дополненного машинно-формируемого ответа 714.

[00113] Следует понимать, что архитектура модели 400 машинного обучения, описанная со ссылкой на фиг. 4, была значительно упрощена для простоты понимания; и фактическая реализация архитектуры 400 модели машинного обучения может включать в себя дополнительные слои и/или блоки, как описано, например, в статье за авторством Vaswani *и др.*, ссылка на которую дана выше. Например, в некоторых реализациях архитектуры 400 модели машинного обучения каждый из заданного блока 404 кодера и заданного блока 405 декодера также может включать в себя операции нормализации слоя. Кроме того, формирование выходных данных 425 может включать в себя применение функции нормализации sof™ax на выходе стека слоев 403 декодера и т.д. Специалисту в данной области техники понятно, что эти операции обычно используются в нейронных сетях и моделях глубокого обучения, таких как архитектура 400 модели машинного обучения.

Модель вложения

[00114] Как упоминалось выше, модель 302 вложения может быть выполнена с возможностью формирования векторных вложений введенного в нее текста, таких как

текстовые представления запросов пользователя, в соответствующем пространстве векторных вложений. В соответствии с некоторыми неограничивающими вариантами осуществления настоящей технологии модель 302 вложения может быть реализована как алгоритм вложения текста, такой как один из алгоритмов 410, 415 входного и целевого вложений, описанных выше.

[00115] Однако в других неограничивающих вариантах осуществления настоящей технологии модель 302 вложения может быть независимой моделью машинного обучения, которую сервер 202 можно сконфигурировать для обучения для формирования соответствующих текстовых векторных вложений. Например, в некоторых неограничивающих вариантах осуществления настоящей технологии модель 302 вложения может быть реализована на основе нейронной сети (NN) с долговременной кратковременной памятью (LS™), как описано, например, в статье за авторствомChang *и др.* под названием *«Expectation-Regulated Neural Model for Event Mention Extraction»*, опубликованной Сингапурским университетом технологий и дизайна в июне 2016 года, содержание которой полностью включено в настоящее описание посредством ссылки.

[00116] В другом примере модель 302 вложения может быть реализована на основе архитектуры 400 модели машинного обучения, описанной выше. В некоторых неограничивающих вариантах осуществления настоящей технологии модель 302 вложения может включать в себя только стек слоев 402 кодера, то есть лишенный каких-либо блоков декодера, имеющий, например, 12, 24 или 36 блоков кодера, реализованных аналогично данному блоку 404 кодера, описанному выше. В этом случае модель 302 вложения может называться моделью представлений двунаправленного кодера от трансформеров (Bidirectional Encoder Representations from Transformers, BERT).

[00117] В этих вариантах осуществления, во-первых, сервер 202 может быть выполнен с возможностью предварительной подготовки модели 302 вложения с использованием цели моделирования маскированного языка (MLM). В частности, цель MLM основана на одной из двух целей неконтролируемого обучения, используемых в BERT, которая используется для изучения текстовых представлений из коллекций немаркированных документов. Эти документы могут включать в себя множество текстовых документов из общественного достояния, таких как десятки тысяч или сотни тысяч, или даже миллионы различных текстовых документов.

[00118] Кроме того, для предварительной подготовки с целью MLM сервер 202 может быть выполнен с возможностью: (i) токенизации текста в каждом из текстовых документов во входных данных 412 для модели 302 вложения, используя, например, предварительно созданный словарь токенов; (ii) с использованием индексов токенов в предварительно созданном словаре токенов, формирования для каждой текстовой единицы, такой как предложение или абзац, соответствующего экземпляра множества входных векторов 414; (ii) маскирования в заданном векторе соответствующего экземпляра множества входных векторов 414 один или более токенов путем замены их специальным токеном [MASK]. В некоторых неограничивающих вариантах осуществления настоящей технологии сервер 202 может быть выполнен с возможностью определения предварительно созданного словаря токенов с использованием схемы кодирования пары байтов WordPiece, используемой в BERT, с достаточно большим размером словаря. Например, в некоторых реализациях размер словаря может составлять приблизительно 120 000 токенов. В некоторых реализациях может выполняться предварительная обработка текста, например - преобразование всех слов в нижний регистр и выполнение нормализации Unicode NFC. Схема кодирования пары байтов WordPiece, которая может использоваться в некоторых реализациях для

формирования словаря токенов, описана, например, в источнике Rico Sennrich *и др.*, «*Neural Machine Translation of Rare Words with Subword Units»,* Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), страницы 1715-1725, 2016.

[00119] Кроме того, сервер 202 может быть выполнен с возможностью подачи множества входных векторов 414, содержащих маскированные токены, в модель 302 вложения, тем самым обучая модель 302 вложения прогнозировать вероятности данного маскированного токена, соответствующего токенам в предварительно созданном словаре токенов. Это делается на основе выходных данных 425 (которые являются вектором) последнего слоя модели 302 вложения, которые соответствуют замаскированным токенам. Более конкретно, во время предварительного обучения сервер 202 может быть выполнен с возможностью оптимизации разницы между фактическими замаскированными токенами (т. е. «наземной истиной») и предсказанными вероятностями. Эта разница может быть выражена функцией потерь (такой как функция кросс-энтропийных потерь, упомянутая выше). Поступая таким образом, сервер 202 может быть выполнен с возможностью корректировки весов в модели 302 вложения для уменьшения потерь.

[00120] Следует четко понимать, что в других неограничивающих вариантах осуществления настоящей технологии модель 302 вложения может быть предварительно обучена, как описано выше, сторонним сервером, к которому серверу 202 может быть дополнительно предоставлен доступ для использования модели 302 вложения.

[00121] Кроме того, сервер 202 может быть выполнен с возможностью использования модели 302 вложения для формирования соответствующих векторных вложений входного текста, таких как соответствующее векторное вложение 306 текстового представления заданного человеческого запроса 212. Более конкретно, сервер 202 может быть выполнен с возможностью формирования соответствующего векторного вложения 306 в качестве выходных данных скрытых блоков кодера стека слоев 402 кодера модели 302 вложения, предварительно обученной, как описано выше. Например, в некоторых неограничивающих вариантах осуществления настоящей технологии сервер 202 может быть выполнен с возможностью формирования соответствующего векторного вложения 306 в качестве объединенных выходных данных предварительно определенного числа (например, шестого, четырех или двух) блоков внешнего кодера модели 302 вложения, каждый из которых может содержать, в определенных реализациях, вектор длиной 768 значений с плавающей запятой, как указано выше.

[00122] Таким образом, модель 302 вложения может быть обучена формировать соответствующие векторные вложения входного текста, так что векторные вложения, связанные с текстовыми единицами, которые сходны по смыслу, были бы ближе друг к другу в соответствующем пространстве векторных вложений; тогда как соответствующие векторные вложения, связанные с текстовыми единицами, которые не близки по смыслу, то есть семантически не похожи, будут дальше друг от друга в соответствующем пространстве вложений.

[00123] Кроме того, сервер 202 может быть выполнен с возможностью использования модели 302 вложения для формирования соответствующих векторных вложений для генеративной модели 304, таких как соответствующее векторное вложение 306 заданного человеческого запроса 212, на основе которого генеративная модель 304 может быть выполнена с возможностью формирования текстового представления соответствующего дополненного машинно-формируемого ответа 714.

Генеративная модель

[00124] Как упоминалось выше, согласно некоторым неограничивающим вариантам осуществления настоящей технологии, генеративная модель 304 системы 210 чат-бота может быть выполнена с возможностью формирования на основе соответствующего векторного вложения 306 текстового представления заданного человеческого запроса 212 текстового представления соответствующего дополненного машинно-формируемого ответа 714, который будет дополнительно предоставлен пользователю 206.

[00125] Соответствующий дополненный машинно-формируемый ответ 714 может указывать, например, на следующую диалоговую строку системы 210 чат-бота в ответ на заданный человеческий запрос 212. Например, если заданный человеческий запрос 212 гласит: *«Эй! Я бы хотел посмотреть хороший мультфильм»*, соответствующий дополненный машинно-формируемый ответ 714 может свидетельствовать о следующей фразе*: «Конечно! Посмотрите «Пиноккио» Гильермо дель Торо. Сейчас он один из самых популярных».*

[00126] В некоторых неограничивающих вариантах осуществления настоящей технологии генеративная модель 304 может быть реализована на основе нейронной сети (NN), такой как LS™ NN или рекуррентная NN. Однако согласно некоторым неограничивающим вариантам осуществления настоящей технологии генеративная модель 304 может быть реализована как модель NN на основе Трансформера. С этой целью генеративная модель 304 может включать в себя некоторые или все компоненты архитектуры 400 модели машинного обучения, описанной выше.

[00127] Например, в некоторых неограничивающих вариантах осуществления настоящей технологии генеративная модель 304 может включать в себя один блок кодера и тринадцать блоков декодера, реализованных аналогично данному блоку 404 кодера и заданному блоку 405 декодера, соответственно, как описано выше. В других неограничивающих вариантах осуществления настоящей технологии генеративная модель 304 может не включать в себя блоки кодера и несколько блоков декодера, таких как, например, 6, 12 или 96, и в этом случае генеративную модель 304 можно назвать моделью генеративного предварительно обученного трансформера (Generative Pre-trained Transformer, GPT). Напротив, в других неограничивающих вариантах осуществления настоящей технологии генеративная модель 304 может включать в себя только блоки кодера, такие как, например, 12, 24 или 36, и не включать в себя блоки декодера, и в этом случае, аналогично модели 302 вложения, генеративная модель 304 может называться моделью BERT.

[00128] Другие конфигурации стека слоев 402 кодера и стека слоев 403 декодера для реализации генеративной модели 304 также предусмотрены, не выходя за рамки настоящей технологии.

[00129] Кроме того, как отмечено выше, вместо по меньшей мере одного из алгоритма 410 входного вложения и алгоритма 415 целевого вложения для формирования входных и целевых множеств входных и целевых векторов 414, 419 для генеративной модели 304, в некоторых неограничивающих вариантах осуществления настоящей технологии сервер 202 может быть выполнен с возможностью использования модели 302 вложения, обученной, как описано выше.

[00130] В целом, согласно некоторым неограничивающим вариантам осуществления настоящей технологии, сервер 202 может быть выполнен с возможностью выполнения двух соответствующих процессов в отношении генеративной модели 304. Первый процесс из двух процессов представляет собой процесс обучения, в котором сервер 202 выполнен с возможностью обучения генеративной модели 304 на основе обучающего набора данных для формирования текстового представления соответствующего

дополненного машинно-формируемого ответа 714, который будет рассмотрен ниже. Второй процесс представляет собой процесс в процессе использования, в котором сервер 202 выполнен с возможностью применения обученной таким образом генеративной модели 304 для ввода текстовых представлений, которые будут описаны в соответствии с некоторыми неограничивающими вариантами осуществления настоящей технологии, приведенными ниже.

Процесс обучения

[00131] В соответствии с некоторыми неограничивающими вариантами осуществления настоящей технологии сервер 202 может быть выполнен с возможностью обучения генеративной модели 304 формированию соответствующих машинно-формируемых ответов в два этапа. Во время первого этапа обучения, также называемого этапом «предобучения», сервер 202 может быть выполнен с возможностью обучения генеративной модели 304 для формирования текстового представления соответствующего дополненного машинно-формируемого ответа 714 на заданный человеческий запрос 212 на основе обучающих диалоговых пар «запрос-ответ» на естественном языке.

[00132] Кроме того, во время второго этапа обучения, следующего за первым этапом, который также называется этапом «точной настройки», наряду с обучающими диалоговыми парами, сервер 202 также может быть выполнен с возможностью использования фактических данных, обеспечивая контекст для соответствующих обучающих диалоговых пар. При этом сервер 202 может быть выполнен с возможностью обучения генеративной модели 304 для формирования текстовых представлений более естественных и разнообразных машинно-формируемых ответов на запросы пользователей.

[00133] Со ссылкой на Фиг. 5, здесь изображена схематическая диаграмма первого этапа обучения сервера 202, генеративной модели 304 для формирования соответствующего дополненного машинно-формируемого ответа 714 на заданный человеческий запрос 212 в соответствии с некоторыми неограничивающими вариантами осуществления настоящей технологии.

[00134] В некоторых неограничивающих вариантах осуществления настоящей технологии сервер 202 может быть выполнен с возможностью размещения (или иным образом иметь удаленный доступ через сеть 208 связи) базу данных диалогов 502, выполненную с возможностью хранения диалоговых данных. В соответствии с некоторыми неограничивающими вариантами осуществления настоящей технологии в диалоговых данных хранятся текстовые представления множества обучающих диалоговых пар, созданных на естественном языке, например, пользователями-людьми. Таким образом, заданная обучающая диалоговая пара из множества обучающих диалоговых пар, хранящихся в базе данных диалогов 502, может включать в себя: (i) текстовое представление заданного обучающего человеческого запроса 503 *(*например, *«Эй, ты уже смотрел новый Аватар? Видеоэффекты просто потрясающие!»); и* (ii) текстовое представление соответствующего обучающего человеческого ответа 505 в ответ на заданный обучающий человеческий запрос 503 (чтение, например: *«Стоп! Не спойлери - сегодня вечером в кино»*).

[00135] Способы конфигурации сервера 202 для заполнения базы данных диалогов 502 не ограничены. Например, в некоторых неограничивающих вариантах осуществления настоящей технологии сервер 202 может быть выполнен с возможностью обхода веб-ресурсов сети 208 связи, включая общедоступные разговоры, которые могут включать в себя, помимо прочего: (i) диалоговые строки из общедоступной литературы;

(ii) диалоговые строки, образованные комментариями пользователей в социальных сетях (таких как социальная сеть VK.COM™); (iii) диалоговые строки, образованные отзывами пользователей, оставленными как минимум на одной из платформ потокового аудио, платформ потокового видео и платформ онлайн-листинга; (iv) диалоговые строки, сформированные дискуссиями на форуме; (v) диалоговые строки из записей телефонных разговоров между различными поставщиками услуг и их клиентами; и тому подобное. Однако в других неограничивающих вариантах осуществления настоящей технологии база данных диалогов 502 может быть предварительно заполнена сторонним сервером (не показан) без отклонения от объема настоящей технологии.

[00136] Таким образом, для обучения генеративной модели 304 на первом этапе обучения сервер 202 может быть выполнен с возможностью формирования первого множества обучающих цифровых объектов, заданный обучающий цифровой объект 504 из которых включает в себя: (i) текстовое представление заданного обучающего человеческого запроса 503; и (ii) соответствующую метку, являющуюся текстовым представлением соответствующего обучающего человеческого ответа 505. Кроме того, сервер 202 может быть выполнен с возможностью подачи каждого из первого множества обучающих цифровых объектов в генеративную модель 304 без введения в нее соответствующих меток обучающих цифровых объектов. Другими словами, при подаче в генеративную модель 304 заданного обучающего цифрового объекта 504, в некоторых неограничивающих вариантах осуществления настоящей технологии, сервер 202 может быть выполнен с возможностью подачи только заданного обучающего человеческого запроса 503 (без подачи на него текстового представления соответствующего обучающего человеческого ответа 505), тем самым вызывая формирование генеративной моделью 304 текстового представления соответствующего обучающего машинно-формируемого ответа 514. Как отмечалось выше, для подачи заданного обучающего человеческого запроса 503 в генеративную модель 304 сервер 202 может быть выполнен с возможностью применения к нему модели 302 вложения, обученной, как описано выше, тем самым формируя соответствующее обучающее векторное вложение 506 для дальнейшей подачи в генеративную модель 304.

[00137] Кроме того, сервер 202 может быть выполнен с возможностью оптимизации разницы между соответствующим обучающим машинно-формируемым ответом 514 и соответствующим обучающим человеческим ответом 505, который может быть выражен функцией потерь, примеры и подходы к оптимизации которых перечислены выше без исчерпывающего перечня. Кроме того, используя алгоритм обратного распространения, на каждой обучающей итерации сервер 202 может быть выполнен с возможностью корректировки весов генеративной модели 304, тем самым предварительно обучая генеративную модель 304 для формирования текстового представления соответствующего дополненного машинно-формируемого ответа 714 на заданный человеческий запрос 212. Кроме того, сервер 202 может быть выполнен с возможностью точной настройки генеративной модели 304 для формирования текстового представления соответствующего дополненного машинно-формируемого ответа 714.

[00138] Со ссылкой на Фиг. 6, здесь изображена схематическая диаграмма второго этапа обучения сервера 202, генеративной модели 304 для формирования соответствующего дополненного машинно-формируемого ответа 714 на заданный человеческий запрос 212 в соответствии с некоторыми неограничивающими вариантами осуществления настоящей технологии.

[00139] Как упоминалось выше, во время второго этапа обучения, согласно некоторым неограничивающим вариантам осуществления настоящей технологии,

наряду с обучающими диалоговыми парами, сервер 202 может быть выполнен с возможностью использования определенных фактических данных. С этой целью, в соответствии с некоторыми неограничивающими вариантами осуществления настоящей технологии, сервер 202 может быть выполнен с возможностью размещения (или иного удаленного доступа) базы данных 602 фактов, выполненной с возможностью хранения данных о фактах, включающих в себя текстовые представления различных фактов. Характер этих фактов не ограничен. Например, на общеизвестные факты могут указывать такие факты, как *«Земля круглая», «COVID-19 - заболевание, вызываемое коронавирусом SARS-CoV-2», «Москва - столица России»* и тому подобное. Однако в некоторых неограничивающих вариантах осуществления настоящей технологии база данных фактов 602 может включать в себя текстовые представления фактов, характерных для определенных отраслей, таких как киноиндустрия, юридическая отрасль, индустрия товаров повседневного спроса (FMCG) и т.п. В связи с этим данная запись в базе данных фактов 602 может иметь следующий вид: *«Аватар-2 заработал более 134 миллионов долларов по всему миру в первые выходные», «Мэрил Стрип выиграла три премии «Оскар», «28 февраля 2023 года шоу «Cunk on Earth» получило рейтинг 8,1 на IMDB»* и тому подобное.

[00140] Способ заполнения базы данных фактов 602 не ограничен. В некоторых неограничивающих вариантах осуществления настоящей технологии сервер 202 может быть выполнен с возможностью заполнения базы данных фактов 602 путем обхода различных ресурсов сети 208 связи. Например, сервер 202 может быть выполнен с возможностью обхода определенных справочных ресурсов (таких как онлайн-энциклопедия Wikipedia™, онлайн-энциклопедия Britannica™ и т.п.), новостных онлайн-порталов (таких как новостной онлайн-портал Yandex.News™, онлайн-новостной портал Rambler™ и т.п.) или справочных и новостных страниц по меньшей мере одной из платформ онлайн-листинга, платформ потокового аудио и платформ потокового видео. Кроме того, в других неограничивающих вариантах осуществления настоящей технологии база данных фактов 602 может быть предварительно заполнена сторонним сервером без сервера 202 и сделана доступной для него, не выходя за рамки настоящей технологии.

[00141] Кроме того, в некоторых неограничивающих вариантах осуществления настоящей технологии база данных фактов 602 может обновляться время от времени, например, регулярно (раз в день, раз в неделю, раз в месяц и т. д.), чтобы факты, хранящиеся в базе данных 602 фактов, были актуальными.

[00142] Кроме того, на основе диалоговых данных из базы данных диалогов 502 и данных о фактах из базы данных фактов 602 сервер 202 может быть выполнен с возможностью формирования второго множества обучающих цифровых объектов, при этом заданный другой обучающий цифровой объект 604 из которых включает в себя: (i) текстовое представление заданного обучающего человеческого запроса 503; (ii) текстовое представление соответствующего обучающего факта 603 из базы данных фактов 602; и (iii) соответствующая метка является текстовым представлением соответствующего обучающего человеческого ответа 505.

[00143] В некоторых неограничивающих вариантах осуществления настоящей технологии второе множество обучающих цифровых объектов может включать в себя обучающие диалоговые пары, которые отличаются от пар, используемых для формирования первого множества обучающих цифровых объектов. В других неограничивающих вариантах осуществления настоящей технологии второе множество обучающих цифровых объектов может включать в себя обучающие диалоговые пары,

которые по меньшей мере частично совпадают с парами, используемыми для формирования первого множества обучающих цифровых объектов. В некоторых неограничивающих вариантах осуществления настоящей технологии количество элементов во втором множестве обучающих цифровых объектов может быть равно количеству элементов в первом множестве обучающих цифровых объектов. В других неограничивающих вариантах осуществления настоящей технологии количество элементов во втором множестве обучающих цифровых объектов может отличаться от первого множества обучающих цифровых объектов, например меньше. Только в качестве примера, в тех вариантах осуществления, где количество элементов в первом множестве обучающих цифровых объектов составляет около 600 миллионов, количество элементов во втором множестве обучающих цифровых объектов может составлять около 30, 40 или даже 200 или 500 тысяч.

[00144] В соответствии с некоторыми неограничивающими вариантами осуществления настоящей технологии сервер 202 может быть выполнен с возможностью идентификации соответствующего обучающего факта 603 как релевантного по меньшей мере одному из заданного обучающего человеческого запроса 503 и соответствующего обучающего человеческого ответа 505. С этой целью в некоторых неограничивающих вариантах осуществления настоящей технологии сервер 202 может быть выполнен с возможностью: (i) подачи текстовых представлений каждой диалоговой пары из базы данных диалогов 502 и текстовые представления фактов из базы данных фактов 602 в модель 302 вложений для формирования соответствующих векторных вложений в соответствующем пространстве векторных вложений; и (ii) определения соответствующего обучающего факта 603 как факта, связанного с соответствующим векторным вложением, которое наиболее близко к соответствующим векторным вложениями по меньшей мере одного из заданного обучающего человеческого запроса 503 и соответствующего обучающего человеческого ответа 505 в соответствующем пространстве векторных вложений.

[00145] Для определения соответствующего векторного вложения, связанного с соответствующим обучающим фактом 603, которое является ближайшим к соответствующим векторным вложениями по меньшей мере одного из заданного обучающего человеческого запроса 503 и соответствующего обучающего человеческого ответа 505, в некоторых неограничивающих вариантах осуществления настоящей технологии сервер 202 может быть выполнен с возможностью применения алгоритма K ближайших соседей, который выполнен с возможностью идентификации числа K ближайших соседей векторных вложений вокруг заданного, например, по меньшей мере одного из заданного обучающего человеческого запроса 503 и соответствующего обучающего человеческого ответа 505. Значение K может быть задано заранее, например, оператором системы 210 чат-бота и может составлять, например, 3, 5 или 10.

[00146] Однако в других неограничивающих вариантах осуществления настоящей технологии, вместо или в дополнение к алгоритму K ближайших соседей, для идентификации соответствующего векторного вложения, связанного с соответствующим обучающим фактом 603, который имеет отношение к соответствующим векторным вложениями по меньшей мере одного из заданного обучающего человеческого запроса 503 и соответствующего обучающего человеческого ответа 505, в некоторых неограничивающих вариантах осуществления настоящей технологии сервер 202 может быть выполнен с возможностью применения эвристического алгоритма, такого как функция ранжирования. Говоря в общем, функция ранжирования выполнена с возможностью определения релевантных поисковых запросов (соответствующий

обучающий факт 603) для соответствующих поисковых запросов (по меньшей мере, одного из заданного обучающего человеческого запроса 503 и соответствующего обучающего человеческого ответа 505) на основе статистики термина «частота - обратная частота документа» (TF-IDF). Например, в некоторых неограничивающих вариантах осуществления настоящей технологии функция ранжирования может содержать одну из функций ранжирования Okapi BM25 и BM25+.

[00147] Кроме того, после идентификации из базы данных фактов 602 соответствующих обучающих фактов для каждой из диалоговых пар во втором множестве обучающих цифровых объектов, как описано выше, сервер 202 может быть выполнен с возможностью подачи каждого из второго множества обучающих цифровых объектов в генеративную модель 304, тем самым точно настраивая генеративную модель 304 для формирования текстовых представлений соответствующих дополненных машинно-формируемых ответов на пользовательские запросы пользователей системы 210 чат-бота. Более конкретно, в этом отношении сервер 202 может быть выполнен с возможностью: (i) для каждого из второго множества обучающих цифровых объектов, таких как заданный другой обучающий цифровой объект 604, формировать объединение текстовых представлений заданного обучающего человеческого запроса 503 и соответствующего обучающего факта 603, идентифицированного с использованием модели 302 вложения; и (ii) подавать объединение текстовых представлений заданного обучающего человеческого запроса 503 и соответствующего обучающего факта 603 в генеративную модель 304, тем самым вызывая формирование генеративной моделью 304 соответствующего другого обучающего машинно-формируемого ответа 614.

[00148] В некоторых неограничивающих вариантах осуществления настоящей технологии сервер 202 может быть выполнен с возможностью формирования объединения текстовых представлений заданного обучающего человеческого запроса 503 и соответствующего обучающего факта 603 путем создания объединения соответствующих векторных вложений, сформированных, например, моделью 302 вложения, как описано выше. Более конкретно, в этих вариантах осуществления сервер 202 может быть выполнен с возможностью: (i) формирования на основе соответствующих векторных вложений заданного обучающего человеческого запроса 503 и соответствующего обучающего факта 603 соответствующего обучающего объединенного векторного вложения 606; и (ii) подачи соответствующего обучающего объединенного векторного вложения 606 в генеративную модель 304, тем самым вызывая формирование генеративной моделью 304 соответствующего другого обучающего машинно-формируемого ответа 614.

[00149] Кроме того, подобно первому этапу обучения, сервер 202 может быть выполнен с возможностью оптимизации разницы между соответствующим другим обучающим машинно-формируемым ответом 614 и соответствующей меткой, которая является соответствующим обучающим человеческим ответом 505, как описано выше. Таким образом, сервер 202 может быть выполнен с возможностью точной настройки генеративной модели 304 для формирования текстовых представлений соответствующих дополненных машинно-формируемых ответов на запросы пользователей с учетом контекстов, обеспечиваемых соответствующими фактами, определенными как относящиеся к запросам пользователей.

[00150] Кроме того, сервер 202 может быть выполнен с возможностью использования генеративной модели 304, предварительно обученной и точно настроенной, как описано выше, для формирования текстовых представлений соответствующих дополненных машинно-формируемых ответов на запросы пользователей, отправленные в систему

210 чат-бота, которые будут описаны непосредственно ниже со ссылкой на фиг. 7.

Процесс использования

[00151] Со ссылкой на фиг. 7, здесь изображена схема используемого процесса для использования сервером 202 как модели 302 вложения, так и генеративной модели 304 для формирования текстового представления соответствующего дополненного машинно-формируемого ответа 714 на заданный человеческий запрос 212, в соответствии с некоторыми неограничивающими вариантами осуществления настоящей технологии.

[00152] В соответствии с некоторыми неограничивающими вариантами осуществления настоящей технологии во время процесса использования сервер 202 может быть выполнен с возможностью: (i) приема текстового представления заданного человеческого запроса 212; (ii) на основе заданного человеческого запроса 212 идентификации в базе данных фактов 602 текстового представления соответствующего используемого факта 703, относящегося к заданному человеческому запросу 212, с использованием соответствующих его векторных вложений, сформированных моделью 302 вложения, как описано выше в отношении соответствующего обучающего факта 603; (iii) формирования объединения текстовых представлений заданного человеческого запроса 212 и соответствующего используемого факта 703; (iv) подавать объединение текстовых представлений заданного человеческого запроса 212 и соответствующего используемого факта 703 в генеративную модель 304, тем самым вызывая формирование генеративной моделью 304 текстового представления соответствующего дополненного машинно-формируемого ответа 714 на заданный человеческий запрос 212.

[00153] Однако в тех вариантах осуществления, где сервер 202 выполнен с возможностью обучения генеративной модели 304 на основе объединения соответствующих векторных вложений, связанных обучающими человеческими запросами и соответствующих обучающих фактов; во время процесса использования в некоторых неограничивающих вариантах осуществления настоящей технологии для формирования объединения текстовых представлений заданного человеческого запроса 212 и соответствующего используемого факта 703 сервер 202 может быть аналогичным образом выполнен с возможностью формирования объединения соответствующих векторных вложений, сформированных, например, моделью 302 вложения, как описано выше. Более конкретно, в этих вариантах осуществления сервер 202 может быть выполнен с возможностью: (i) подачи текстовых представлений каждого из заданного человеческого запроса 212 и соответствующего используемого факта 703 в модель 302 вложения для формирования их соответствующих векторных вложений; (ii) на основе соответствующих векторных вложений заданного человеческого запроса 212 и соответствующего используемого факта 703 формирования соответствующего используемого объединенного векторного вложения 706; и (iii) подачи соответствующего используемого объединенного векторного вложения 706 в генеративную модель 304, тем самым вызывая формирование генеративной моделью 304 текстового представления соответствующего дополненного машинно-формируемого ответа 714 на заданный человеческий запрос 212.

[00154] Кроме того, сервер 202 может быть выполнен с возможностью передачи текстового представления соответствующего дополненного машинно-формируемого ответа 714 на электронное устройство 204 пользователя 206 для представления ему соответствующего дополненного машинно-формируемого ответа 714. Как упоминалось выше, в тех вариантах осуществления, где система 210 чат-бота дополнительно включает в себя модель TTS, система 210 чат-бота может быть выполнена с возможностью формирования на основе текстового представления соответствующего дополненного

машинно-формируемого ответа 714 соответствующего машинного высказывания для дальнейшего его воспроизведения для пользователя 206 с использованием динамика электронного устройства 204.

[00155] Как упоминалось выше, при рассмотрении контекста соответствующего используемого факта 703, относящегося к заданному человеческому запросу 212, генеративная модель 304 выполнена с возможность формирования соответствующего дополненного машинно-формируемого ответа 714, который может восприниматься пользователем 206 как более естественный и неограниченный, а также относящийся к настоящему, в отличие от соответствующего машинно-формируемого ответа 214, сформированного исключительно на основе заданного человеческого запроса 212. Это может улучшить взаимодействие пользователя 206 с системой 210 чат-бота.

Способ

[00156] Учитывая архитектуру и приведенные выше примеры, можно реализовать способ обучения системы 210 чат-бота формировать машинно-формируемые ответы на запросы пользователей системы 210 чат-бота, такие как соответствующий дополненный машинно-формируемый ответ 714 в ответ на заданный человеческий запрос 212 пользователя 206.

[00157] Как упоминалось выше, в соответствии с некоторыми неограничивающими вариантами осуществления настоящей технологии система 210 чат-бота может содержать: (i) модель 302 вложения, выполненная с возможностью формирования векторных вложений текстовых представлений пользовательских запросов, таких как соответствующее используемое объединенное векторное вложение 706 заданного человеческого запроса 212; и (ii) генеративную модель 304, выполненную с возможностью формирования на основе соответствующего используемого объединенного векторного вложения 706 текстового представления соответствующего дополненного машинно-формируемого ответа 714.

[00158] В соответствии с некоторыми неограничивающими вариантами осуществления настоящей технологии каждая модель 302 вложения и генеративная модель 304 могут быть реализованы как модель машинного обучения на основе Трансформера, имеющая архитектуру 400 модели машинного обучения, подробно описанную выше со ссылкой на фиг. 4.

[00159] Теперь со ссылкой на фиг. 8 изображена блок-схема способа 800 в соответствии с неограничивающими вариантами осуществления настоящей технологии. Способ 800 может быть исполнен сервером 202.

ЭТАП 802: ПОЛУЧЕНИЕ ПРОЦЕССОРОМ ДИАЛОГОВЫХ ДАННЫХ, ВКЛЮЧАЮЩИХ В СЕБЯ (I) ТЕКСТОВЫЕ ПРЕДСТАВЛЕНИЯ ЧЕЛОВЕЧЕСКИХ ЗАПРОСОВ ДИАЛОГОВ ПОЛЬЗОВАТЕЛЕЙ НА ЕСТЕСТВЕННОМ ЯЗЫКЕ; И (II) ТЕКСТОВЫЕ ПРЕДСТАВЛЕНИЯ СООТВЕТСТВУЮЩИХ ЧЕЛОВЕЧЕСКИХ ОТВЕТОВ В ДИАЛОГАХ, ОТВЕЧАЮЩИХ НА ЧЕЛОВЕЧЕСКИЕ ЗАПРОСЫ.

[00160] Как упоминалось выше, в соответствии с некоторыми неограничивающими вариантами осуществления настоящей технологии обучение системы 210 чат-бота формированию соответствующего дополненного машинно-формируемого ответа 714 в ответ на заданный человеческий запрос 212 может включать в себя два этапа обучения. На первом этапе обучения, также называемом этапом «предобучения», сервер 202 может быть выполнен с возможностью обучения генеративной модели 304 системы 210 чат-бота для формирования текстового представления соответствующего дополненного машинно-формируемого ответа 714 на заданный человеческий запрос 212 на основе обучающих диалоговых пар «запрос-ответ» на естественном языке.

[00161] Кроме того, во время второго этапа обучения, следующего за первым этапом, который называется этапом «точной настройки», наряду с диалоговыми парами обучения сервер 202 также может быть выполнен с возможностью использования фактических данных, обеспечивая контекст для соответствующих диалоговых пар обучения. Таким образом, сервер 202 может быть выполнен с возможностью обучения генеративной модели 304 для формирования текстовых представлений более естественных и разнообразных машинно-формируемых ответов на запросы пользователей.

[00162] Таким образом, способ 800 начинается с этапа 802, когда сервер 202 выполнен с возможностью получения, например, из базы данных диалогов 502 диалоговых данных, включающих в себя текстовые представления множества обучающих диалоговых пар на естественном языке. Например, как было описано выше со ссылкой на фиг. 5, данная обучающая диалоговая пара из множества обучающих диалоговых пар, хранящихся в базе данных диалогов 502, может включать в себя: (i) текстовое представление заданного обучающего человеческого запроса 503; и (ii) текстовое представление соответствующего обучающего человеческого ответа 505 в ответ на заданный обучающий человеческий запрос 503.

[00163] Кроме того, как подробно описано со ссылкой на фиг. 5, сервер 202 может быть выполнен с возможностью формирования первого множества обучающих цифровых объектов, заданный обучающий цифровой объект 504 из которых включает в себя: (i) текстовое представление заданного обучающего человеческого запроса 503; и (ii) соответствующая метка является текстовым представлением соответствующего обучающего человеческого ответа 505. Кроме того, сервер 202 может быть выполнен с возможностью подачи каждого из первого множества обучающих цифровых объектов в генеративную модель 304 без введения в нее соответствующих меток обучающих цифровых объектов. Другими словами, при подаче в генеративную модель 304 заданного обучающего цифрового объекта 504, в некоторых неограничивающих вариантах осуществления настоящей технологии, сервер 202 может быть выполнен с возможностью подачи только заданного обучающего человеческого запроса 503 (без подачи на него текстового представления соответствующего обучающего человеческого ответа 505), тем самым вызывая формирование генеративной моделью 304 текстового представления соответствующего обучающего машинно-формируемого ответа 514. Как отмечалось выше, для подачи заданного обучающего человеческого запроса 503 в генеративную модель 304 сервер 202 может быть выполнен с возможностью применения к нему модели 302 вложения, обученной, как описано выше, тем самым формируя соответствующее обучающее векторное вложение 506 для дальнейшей подачи в генеративную модель 304.

[00164] Кроме того, сервер 202 может быть выполнен с возможностью оптимизации разницы между соответствующим обучающим машинно-формируемым ответом 514 и соответствующим обучающим человеческим ответом 505, как описано выше, тем самым предварительно обучая генеративную модель 304 для формирования текстового представления соответствующего дополненного машинно-формируемого ответа 714 на заданный человеческий запрос 212.

[00165] Кроме того, сервер 202 может быть выполнен с возможностью точной настройки генеративной модели 304 для формирования текстового представления соответствующего дополненного машинно-формируемого ответа 714, как будет описано на следующих этапах способа 800.

[00166] Таким образом, способ 800 переходит к этапу 804.

ЭТАП 804: ПОЛУЧЕНИЕ ПРОЦЕССОРОМ ДАННЫХ О ФАКТАХ, ВКЛЮЧАЮЩИХ В СЕБЯ ТЕКСТОВОЕ ПРЕДСТАВЛЕНИЕ ФАКТОВ.

[00167] На этапе 804 сервер 202 может быть выполнен с возможностью получения, например, из базы данных фактов 602 сервера 202 данных о фактах, включающих в себя текстовые представления различных фактов.

[00168] Например, некоторые факты, хранящиеся в базе данных фактов 602, могут указывать на общеизвестность, например, *«Земля круглая», «COVID-19 - это заболевание, вызываемое коронавирусом SARS-CoV-2», «Москва* - столица России» и тому подобное. Однако в некоторых неограничивающих вариантах осуществления настоящей технологии база данных фактов 602 может включать в себя текстовые представления фактов, характерных для определенных отраслей, таких как киноиндустрия, юридическая отрасль, индустрия товаров повседневного спроса (FMCG) и т.п. В связи с этим данная запись в базе данных фактов 602 может иметь следующий вид: *«Аватар-2 заработал более 134 миллионов долларов по всему миру в первые выходные», «Мэрил Стрип выиграла три премии «Оскар», «28 февраля 2023 года шоу «Cunk on Earth» получило рейтинг 8,1 на IMDB»* и тому подобное.

[00169] Способ заполнения базы данных фактов 602 не ограничен. В некоторых неограничивающих вариантах осуществления настоящей технологии сервер 202 может быть выполнен с возможностью заполнения базы данных фактов 602 путем обхода различных ресурсов сети 208 связи. Например, сервер 202 может быть выполнен с возможностью обхода определенных справочных ресурсов (таких как онлайн-энциклопедия Wikipedia™, онлайн-энциклопедия Britannica™ и т.п.), новостных онлайн-порталов (таких как новостной онлайн-портал Yandex.News™, онлайновый новостной портал Rambler™ и т.п.) или справочных и новостных страниц по меньшей мере одной из платформ онлайн-листинга, платформ потокового аудио и платформ потокового видео. Кроме того, в других неограничивающих вариантах осуществления настоящей технологии база данных фактов 602 может быть предварительно заполнена сторонним сервером без сервера 202 и сделана доступной для него, не выходя за рамки настоящей технологии.

[00170] Кроме того, в некоторых неограничивающих вариантах осуществления настоящей технологии база данных фактов 602 может обновляться время от времени, например, регулярно (раз в день, раз в неделю, раз в месяц и т. д.), чтобы факты, хранящиеся в базе данных фактов 602, были актуальными.

[00171] Таким образом, способ 800 переходит к этапу 806.

ЭТАП 806: ИДЕНТИФИКАЦИЯ ПРОЦЕССОРОМ С ИСПОЛЬЗОВАНИЕМ ML-МОДЕЛИ СЕМАНТИЧЕСКОГО ПОДОБИЯ ИЗ ДАННЫХ О ФАКТАХ ДЛЯ ЗАДАННОЙ ДИАЛОГОВОЙ ПАРЫ ВКЛЮЧАЮЩЕЙ В СЕБЯ ТЕКСТОВЫЕ ПРЕДСТАВЛЕНИЯ ЗАДАННОГО ЧЕЛОВЕЧЕСКОГО ЗАПРОСА И СООТВЕТСТВУЮЩЕГО ЧЕЛОВЕЧЕСКОГО ОТВЕТА НА НЕГО ИЗ ДИАЛОГОВЫХ ДАННЫХ, ТЕКСТОВОГО ПРЕДСТАВЛЕНИЯ СООТВЕТСТВУЮЩЕГО ФАКТА, КОТОРОЕ ОТНОСИТСЯ К ПО МЕНЬШЕЙ МЕРЕ ОДНОМУ ИЗ ЗАДАННОГО ЧЕЛОВЕЧЕСКОГО ЗАПРОСА И СООТВЕТСТВУЮЩЕГО ЧЕЛОВЕЧЕСКОГО ОТВЕТА.

[00172] На этапе 806 сервер 202 может быть выполнен с возможностью идентификации в базе данных фактов 602 для заданной диалоговой пары обучения из множества диалоговых пар из базы данных 502 диалогов, соответствующего обучающего факта 603.

[00173] Как подробно описано выше со ссылкой на фиг. 6, в соответствии с

некоторыми неограничивающими вариантами осуществления настоящей технологии сервер 202 может быть выполнен с возможностью идентификации соответствующего обучающего факта 603 как релевантного по меньшей мере одному из заданного обучающего человеческого запроса 503 и соответствующего обучающего человеческого ответа 505. С этой целью в некоторых неограничивающих вариантах осуществления настоящей технологии сервер 202 может быть выполнен с возможностью: (i) подачи текстовых представлений каждой диалоговой пары из базы данных диалогов 502 и текстовых представлений фактов из базы данных фактов 602 в модель 302 вложений для формирования соответствующих векторных вложений в соответствующем пространстве векторных вложений; и (ii) определения соответствующего обучающего факта 603 как факта, связанного с соответствующим векторным вложением, которое наиболее близко к соответствующим векторным вложениями по меньшей мере одного из заданного обучающего человеческого запроса 503 и соответствующего обучающего человеческого ответа 505 в соответствующем пространстве векторных вложений.

[00174] Для определения соответствующего векторного вложения, связанного с соответствующим обучающим фактом 603, которое является ближайшим к соответствующим векторным вложениями по меньшей мере одного из заданного обучающего человеческого запроса 503 и соответствующего обучающего человеческого ответа 505, в некоторых неограничивающих вариантах осуществления настоящей технологии сервер 202 может быть выполнен с возможностью применения алгоритма K ближайших соседей. В других неограничивающих вариантах осуществления настоящей технологии, вместо или в дополнение к алгоритму K ближайших соседей, для идентификации соответствующего векторного вложения, связанного с соответствующим обучающим фактом 603, сервер 202 может быть выполнен с возможностью применения функции ранжирования, такой как одна из функций ранжирования Okapi BM25 и BM25+.

[00175] Таким образом, способ 800 переходит к этапу 808.

ЭТАП 808: ФОРМИРОВАНИЕ ПРОЦЕССОРОМ ОБУЧАЮЩЕГО НАБОРА ДАННЫХ, ВКЛЮЧАЮЩЕГО В СЕБЯ МНОЖЕСТВО ОБУЧАЮЩИХ ЦИФРОВЫХ ОБЪЕКТОВ, КАЖДЫЙ ИЗ КОТОРЫХ ВКЛЮЧАЕТ В СЕБЯ: (I) ТЕКСТОВОЕ ПРЕДСТАВЛЕНИЕ ЗАДАННОГО ЧЕЛОВЕЧЕСКОГО ЗАПРОСА; (II) ТЕКСТОВОЕ ПРЕДСТАВЛЕНИЕ СООТВЕТСТВУЮЩЕГО ФАКТА; И (III) СООТВЕТСТВУЮЩУЮ МЕТКУ, ЯВЛЯЮЩУЮСЯ ТЕКСТОВЫМ ПРЕДСТАВЛЕНИЕМ СООТВЕТСТВУЮЩЕГО ЧЕЛОВЕЧЕСКОГО ОТВЕТА НА ЗАДАННЫЙ ЧЕЛОВЕЧЕСКИЙ ЗАПРОС.

[00176] На этапе 808 сервер 202 может быть выполнен с возможностью формирования второго множества обучающих цифровых объектов, данный другой обучающий цифровой объект 604 из которых включает в себя: (i) текстовое представление заданного обучающего человеческого запроса 503; (ii) текстовое представление соответствующего обучающего факта 603 из базы данных фактов 602; и (iii) соответствующая метка является текстовым представлением соответствующего обучающего человеческого ответа 505.

[00177] Кроме того, используя второе множество обучающих цифровых объектов, сервер 202 может быть выполнен с возможностью точной настройки генеративной модели 304 для формирования текстового представления соответствующего дополненного машинно-формируемого ответа 714 на заданный человеческий запрос 212.

[00178] Таким образом, способ 800 переходит к этапу 810.

ЭТАП 810: ПОДАЧА ПРОЦЕССОРОМ ОБУЧАЮЩЕГО НАБОРА ДАННЫХ В

СИСТЕМУ ЧАТ-БОТА, ПРИ ЭТОМ УПОМЯНУТАЯ ПОДАЧА ВКЛЮЧАЕТ В СЕБЯ: ДЛЯ ЗАДАННОГО ОБУЧАЮЩЕГО ЦИФРОВОГО ОБЪЕКТА ИЗ МНОЖЕСТВА ОБУЧАЮЩИХ ЦИФРОВЫХ ОБЪЕКТОВ, ПОДАЧА ПРОЦЕССОРОМ ОБЪЕДИНЕНИЯ ТЕКСТОВОГО ПРЕДСТАВЛЕНИЯ ЗАДАННОГО ЧЕЛОВЕЧЕСКОГО ЗАПРОСА И ТЕКСТОВОГО ПРЕДСТАВЛЕНИЯ СООТВЕТСТВУЮЩЕГО ФАКТА В ГЕНЕРАТИВНОЙ ML-МОДЕЛИ ДЛЯ ФОРМИРОВАНИЯ ТЕКСТОВОГО ПРЕДСТАВЛЕНИЯ СООТВЕТСТВУЮЩЕГО МАШИННО-ФОРМИРУЕМОГО ОТВЕТА НА ЗАДАННЫЙ ЧЕЛОВЕЧЕСКИЙ ЗАПРОС С УЧЕТОМ КОНТЕКСТА СООТВЕТСТВУЮЩЕГО ФАКТА.

[00179] На этапе 810 сервер 202 может быть выполнен с возможностью подачи каждого из второго множества обучающих цифровых объектов в генеративную модель 304, тем самым точно настраивая генеративную модель 304 для формирования текстовых представлений соответствующих дополненных машинно-формируемых ответов на пользовательские запросы пользователей системы 210 чат-бота. Более конкретно, в некоторых неограничивающих вариантах осуществления настоящей технологии сервер 202 может быть выполнен с возможностью: (i) для каждого из второго множества обучающих цифровых объектов, таких как заданный другой обучающий цифровой объект 604, формирования объединения текстовых представлений заданного обучающего человеческого запроса 503 и соответствующего обучающего факта 603, идентифицированного с использованием модели 302 вложения, как описано выше; и (ii) подачи объединения текстовых представлений заданного обучающего человеческого запроса 503 и соответствующего обучающего факта 603 в генеративную модель 304, тем самым вызывая формирование генеративной моделью 304 соответствующего другого обучающего машинно-формируемого ответа 614.

[00180] В некоторых неограничивающих вариантах осуществления настоящей технологии сервер 202 может быть выполнен с возможностью формирования объединения текстовых представлений заданного обучающего человеческого запроса 503 и соответствующего обучающего факта 603 путем создания объединения соответствующих векторных вложений, сформированных, например, моделью 302 вложения, как описано выше. С этой целью сервер 202 может быть выполнен с возможностью: (i) формирования на основе соответствующих векторных вложений заданного обучающего человеческого запроса 503 и соответствующего обучающего факта 603 соответствующего обучающего объединенного векторного вложения 606; и (ii) подачи соответствующего обучающего объеденного векторного вложения 606 в генеративную модель 304, тем самым вызывая формирование генеративной моделью 304 соответствующего другого обучающего машинно-формируемого ответа 614.

[00181] Таким образом, способ 800 переходит к этапу 812.

ЭТАП 812: ОПТИМИЗАЦИЯ ПРОЦЕССОРОМ РАЗНИЦЫ МЕЖДУ СООТВЕТСТВУЮЩИМ МАШИННО-ФОРМИРУЕМЫМ ОТВЕТОМ И СООТВЕТСТВУЮЩИМ ЧЕЛОВЕЧЕСКИМ ОТВЕТОМ НА ЗАДАННЫЙ ЧЕЛОВЕЧЕСКИЙ ЗАПРОС, ЗА СЧЕТ ЧЕГО ПРОИСХОДИТ ОБУЧЕНИЕ СИСТЕМЫ ЧАТ-БОТА ФОРМИРОВАТЬ МАШИННО-ФОРМИРУЕМЫЕ ОТВЕТЫ НА ЗАПРОСЫ ПОЛЬЗОВАТЕЛЕЙ.

[00182] На этапе 812 сервер 202 может быть выполнен с возможностью оптимизации разницы между соответствующим другим обучающим машинно-формируемым ответом 614 и соответствующей меткой, которая является соответствующим обучающим человеческим ответом 505, который может быть выражен функцией потерь, примеры и подходы к оптимизации которых перечислены выше без исчерпывающего перечня.

Кроме того, используя алгоритм обратного распространения, на каждой итерации обучения сервер 202 можно настроить для корректировки весов генеративной модели 304. Таким образом, сервер 202 может быть выполнен с возможностью точной настройки генеративной модели 304 для формирования текстовых представлений соответствующих дополненных машинно-формируемых ответов на запросы пользователей с учетом контекстов, обеспечиваемых соответствующими фактами, определенными как относящиеся к запросам пользователей.

[00183] Кроме того, как подробно описано со ссылкой на фиг. 7 выше, в соответствии с некоторыми неограничивающими вариантами осуществления настоящей технологии сервер 202 может быть выполнен с возможностью использования системы 210 чат-бота для формирования на основе текстовых представлений заданного человеческого запроса 212 и соответствующего используемого факта 703, текстового представления соответствующего дополненного машинно-формируемого ответа 714.

[00184] Кроме того, сервер 202 может быть выполнен с возможностью передачи текстового представления соответствующего дополненного машинно-формируемого ответа 714 на электронное устройство 204 пользователя 206 для представления ему соответствующего дополненного машинно-формируемого ответа 714. Как упоминалось выше, в тех вариантах осуществления, где система 210 чат-бота дополнительно включает в себя модель TTS, система 210 чат-бота может быть выполнена с возможностью формирования на основе текстового представления соответствующего дополненного машинно-формируемого ответа 714 соответствующего машинного высказывания для дальнейшего его воспроизведения для пользователя 206 с использованием динамика электронного устройства 204.

[00185] Таким образом, способ 800 завершается.

[00186] Таким образом, при рассмотрении контекста соответствующих фактов, определяемых как относящиеся к запросам пользователей, определенные варианты осуществления способа 800 могут позволить формировать машинно-формируемые ответы, которые могут восприниматься пользователями системы 210 чат-бота как более естественные и неограниченные, а также соответствующие настоящему. Это может помочь улучшить взаимодействие пользователей с системой 210 чат-бота.

[00187] Следует четко понимать, что не все технические эффекты, упомянутые здесь, должны быть реализованы в каждом и каждом варианте осуществления настоящей технологии.

[00188] Модификации и улучшения вышеописанных реализаций настоящей технологии могут стать очевидными для специалистов в данной области техники. Предшествующее описание предназначено для того, чтобы быть примерным, а не ограничивающим. Поэтому предполагается, что объем настоящей технологии ограничен лишь объемом прилагаемой формулы изобретения.

(57) Формула изобретения

1. Реализуемый компьютером способ обучения системы чат-бота формировать машинно-формируемые ответы на запросы пользователей системы чат-бота, при этом система чат-бота включает в себя: (i) модель машинного обучения (ML) семантического подобия, выполненную с возможностью идентификации соответствующих фактов, имеющих отношение к запросам пользователей; и (ii) генеративную ML-модель, подлежащую обучению для формирования текстовых представлений машинно-формируемых ответов на запросы пользователей на основе запросов пользователей и соответствующих фактов; при этом способ выполняется сервером, включающим в себя

процессор, при этом способ содержит:

получение процессором диалоговых данных, включающих в себя (i) текстовые представления человеческих запросов диалогов пользователей на естественном языке; и (ii) текстовые представления соответствующих человеческих ответов диалогов, отвечающих на человеческие запросы;

получение процессором данных о фактах, включающих в себя текстовое представление фактов;

идентификацию процессором с использованием ML-модели семантического подобия из данных о фактах для заданной диалоговой пары, включающей в себя текстовые представления заданного человеческого запроса и соответствующего человеческого ответа на него из диалоговых данных, текстового представления соответствующего факта, которое относится к по меньшей мере одному из заданного человеческого запроса и соответствующего человеческого ответа;

формирование процессором обучающего набора данных, включающего в себя множество обучающих цифровых объектов, каждый из которых включает в себя: (i) текстовое представление заданного человеческого запроса; (ii) текстовое представление соответствующего факта; и (iii) соответствующую метку, являющуюся текстовым представлением соответствующего человеческого ответа на заданный человеческий запрос;

подачу процессором обучающего набора данных в систему чат-бота, при этом упомянутая подача включает в себя:

для заданного обучающего цифрового объекта из множества обучающих цифровых объектов подачу процессором объединения текстового представления заданного человеческого запроса и текстового представления соответствующего факта в генеративную ML-модель для формирования текстового представления соответствующего машинно-формируемого ответа на заданный человеческий запрос с учетом контекста соответствующего факта; и

оптимизацию процессором разницы между соответствующим машинно-формируемым ответом и соответствующим человеческим ответом на заданный человеческий запрос, за счет чего происходит обучение системы чат-бота формировать машинно-формируемые ответы на запросы пользователей.

2. Способ по п. 1, в котором идентификация с использованием ML-модели семантического подобия соответствующего факта для заданной диалоговой пары текстовых представлений содержит:

подачу процессором каждого из: (i) текстовых представлений человеческих запросов; (ii) текстовых представлений человеческих ответов; и (iii) текстовых представлений фактов в ML-модель семантического подобия для формирования соответствующих векторных вложений каждого из (i) текстовых представлений человеческих запросов; (ii) текстовых представлений человеческих ответов; и (iii) текстовых представлений фактов;

отображение процессором соответствующих векторных вложений в векторное пространство; и

идентификацию процессором соответствующего факта, относящегося к по меньшей мере одному из заданного человеческого запроса и соответствующего человеческого ответа, как факта, связанного с соответствующим векторным вложением, которое наиболее близко к соответствующему векторному вложению по меньшей мере одного из заданного человеческого запроса и соответствующего человеческого ответа в векторном пространстве.

3. Способ по п. 2, в котором идентификация дополнительно содержит применение процессором алгоритма k-ближайших соседей.

4. Способ по п. 2, в котором идентификация дополнительно содержит применение процессором эвристического алгоритма.

5. Способ по п. 4, в котором эвристический алгоритм содержит функцию ранжирования.

6. Способ по п. 5, в котором функция ранжирования представляет собой функцию ранжирования BM25.

7. Способ по п. 1, в котором соответствующий факт предназначен для предоставления контекста заданной диалоговой паре.

8. Способ по п. 1, в котором объединение текстового представления заданного человеческого запроса и текстового представления соответствующего факта содержит объединение соответствующих векторных вложений заданного человеческого запроса и соответствующего факта в заданном векторном пространстве.

9. Способ по п. 1, в котором оптимизация разницы содержит оптимизацию функции потерь, представляющей разницу между соответствующим машинно-формируемым ответом и соответствующим человеческим ответом.

10. Способ по п. 1, дополнительно содержащий использование системы чат-бота для формирования машинно-формируемых ответов на запросы пользователей, при этом использование содержит:

получение процессором текстового представления используемого человеческого запроса заданного пользователя;

идентификацию процессором с использованием ML-модели семантического подобия из данных о фактах соответствующего факта, относящегося к используемому человеческому запросу;

подачу процессором объединения текстовых представлений используемого человеческого запроса и соответствующего факта в генеративную ML-модель, тем самым вызывая формирование генеративной ML-моделью текстового представления соответствующего используемого машинно-формируемого ответа в ответ на используемый человеческий запрос с учетом контекста соответствующего факта.

11. Способ по п. 1, в котором каждая из ML-модели семантического подобия и генеративной ML-модели представляет собой ML-модель на основе Трансформера.

12. Способ по п. 11, в котором генеративная ML-модель лишена части кодера ML-модели на основе Трансформера.

13. Способ по п. 1, в котором система чат-бота представляет собой одно из следующего: (i) система чат-бота для преобразования текста в текст; (ii) система чат-бота для преобразования текста в речь; (iii) система чат-бота для преобразования речи в текст; и (iv) система чат-бота для преобразования речи в речь.

14. Сервер для обучения системы чат-бота формировать машинно-формируемые ответы на запросы пользователей системы чат-бота, при этом система чат-бота включает в себя: (i) модель машинного обучения (ML) семантического подобия, выполненную с возможностью идентификации соответствующих фактов, имеющих отношение к запросам пользователей; и (ii) генеративную ML-модель, подлежащую обучению для формирования текстовых представлений машинно-формируемых ответов на запросы пользователей на основе запросов пользователей и соответствующих фактов; сервер, содержащий: (i) процессор и (ii) не-кратковременный считываемый компьютером носитель, хранящий инструкции;

при этом процессор, при исполнении инструкций, выполнен с возможностью:

получения диалоговых данных, включающих в себя (i) текстовые представления человеческих запросов из диалогов пользователей на естественном языке; и (ii) текстовые представления соответствующих человеческих ответов из диалогов, отвечающих на человеческие запросы;

получения данных о фактах, включающих в себя текстовые представления фактов;

идентификации из данных о фактах, используя ML-модель семантического подобия, для заданной диалоговой пары, включающей в себя текстовые представления заданного человеческого запроса и соответствующий человеческий ответ, отвечающий на него, из диалоговых данных, текстового представления соответствующего факта, относящегося к по меньшей мере одному из заданного человеческого запроса и соответствующего человеческого ответа;

формирования обучающего набора данных, включающего в себя множество обучающих цифровых объектов, каждый из которых включает в себя: (i) текстовое представление заданного человеческого запроса; (ii) текстовое представление соответствующего факта; и (iii) соответствующую метку, являющуюся текстовым представлением соответствующего человеческого ответа на заданный человеческий запрос;

подачи обучающего набора данных в систему чат-бота посредством:

для заданного обучающего цифрового объекта из множества обучающих цифровых объектов, подачи процессором объединения текстового представления заданного человеческого запроса и текстового представления соответствующего факта в генеративную ML-модель для формирования текстового представления соответствующего машинно-формируемого ответа на заданный человеческий запрос с учетом контекста соответствующего факта; и

оптимизации разницы между соответствующим машинно-формируемым ответом и соответствующим человеческим ответом на заданный человеческий запрос, за счет чего происходит обучение системы чат-бота формировать машинно-формируемые ответы на запросы пользователей.

15. Сервер по п. 14, в котором для идентификации соответствующего факта для заданной диалоговой пары текстовых представлений с использованием ML-модели семантического подобия процессор выполнен с возможностью:

подачи каждого из: (i) текстовых представлений человеческих запросов; (ii) текстовых представлений человеческих ответов; и (iii) текстовых представлений фактов в ML-модель семантического подобия для формирования соответствующих векторных вложений каждого из (i) текстовых представлений человеческих запросов; (ii) текстовых представлений человеческих ответов; и (iii) текстовых представлений фактов;

отображать соответствующие векторные вложения в векторное пространство; и

идентифицировать соответствующий факт, относящийся к по меньшей мере одному из заданного человеческого запроса и соответствующего человеческого ответа, как факт, связанный с соответствующим векторным вложением, которое наиболее близко к соответствующему векторному вложению по меньшей мере одного из заданного человеческого запроса и соответствующего человеческого ответа в векторном пространстве.

16. Сервер по п. 14, в котором для идентификации соответствующего факта процессор дополнительно выполнен с возможностью применения алгоритма k-ближайших соседей.

17. Сервер по п. 14, в котором для оптимизации разницы процессор выполнен с возможностью оптимизации функции потерь, представляющей разницу между соответствующим машинно-формируемым ответом и соответствующим человеческим

ответом.

18. Сервер по п. 14, в котором процессор дополнительно выполнен с возможностью использования системы чат-бота для формирования машинно-формируемых ответов на запросы пользователей посредством:

получения текстового представления используемого человеческого запроса заданного пользователя;

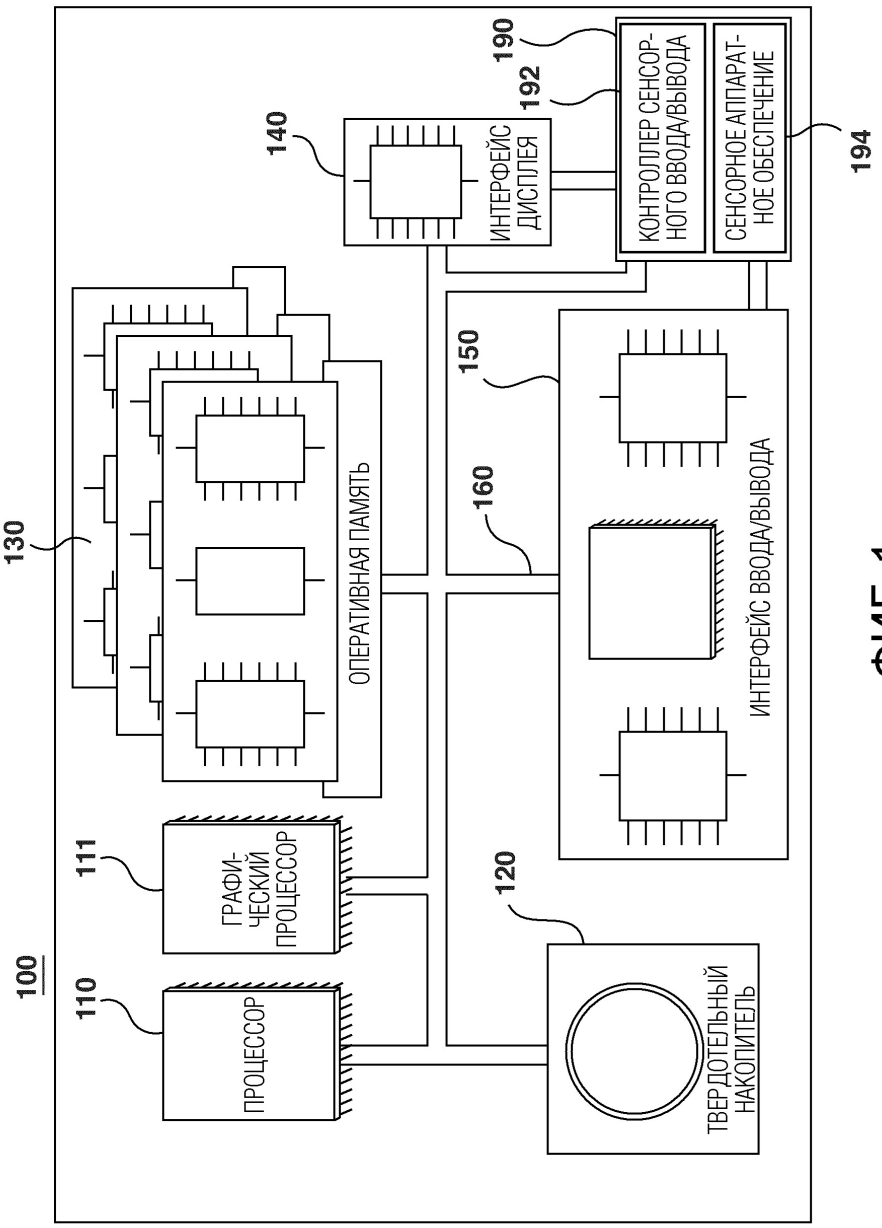
идентификации с использованием ML-модели семантического подобия из данных о фактах соответствующего факта, относящегося к используемому человеческому запросу;

подачи объединения текстовых представлений используемого человеческого запроса и соответствующего факта в генеративную ML-модель, тем самым вызывая формирование генеративной ML-моделью текстового представления соответствующего используемого машинно-формируемого ответа в ответ на используемый человеческий запрос с учетом контекста соответствующего факта.

19. Сервер по п. 14, в котором каждая из ML-модели семантического подобия и генеративной ML-модели представляет собой ML-модель на основе Трансформера.

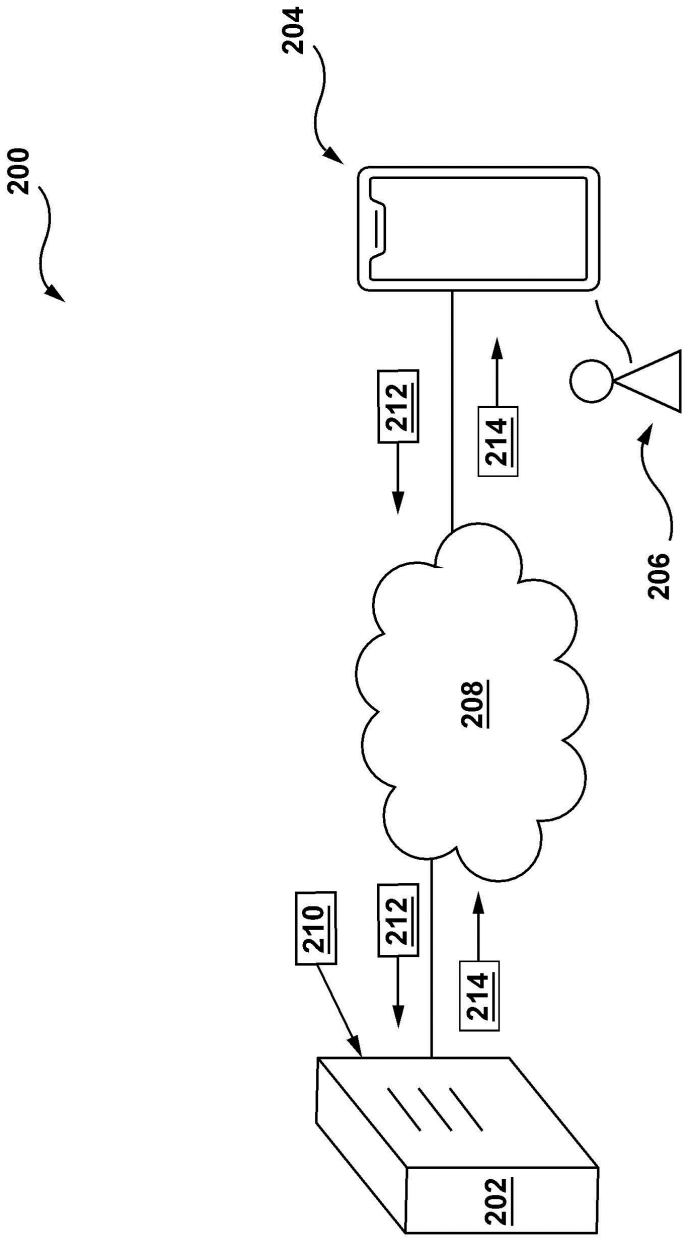
20. Сервер по п. 19, в котором генеративная ML-модель лишена части кодера ML-модели на основе Трансформера.

1


100
110
ПРОЦЕССОР
111
ГРАФИЧЕСКИЙ ПРОЦЕССОР
120
ТВЕРДОТЕЛЬНЫЙ НАКОПИТЕЛЬ
130
ОПЕРАТИВНАЯ ПАМЯТЬ
140
ИНТЕРФЕЙС ДИСПЛЕЯ
150
160
ИНТЕРФЕЙС ВВОДА/ВЫВОДА
190
КОНТРОЛЛЕР СЕНСОРНОГО ВВОДА/ВЫВОДА
192
194
СЕНСОРНОЕ АППАРАТНОЕ ОБЕСПЕЧЕНИЕ

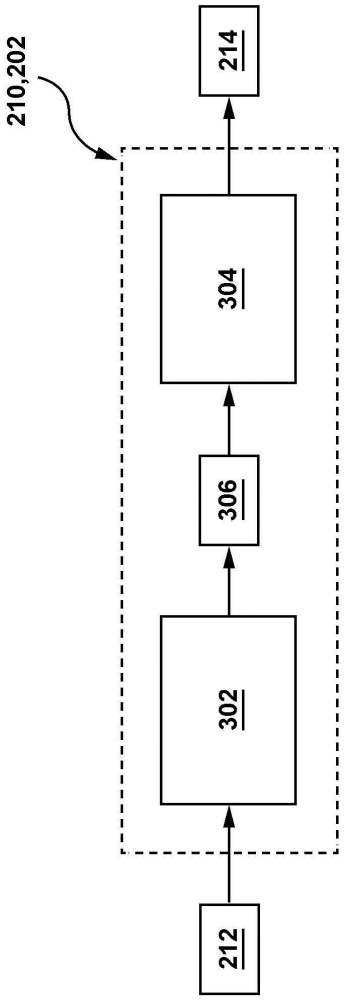
ФИГ. 1

2



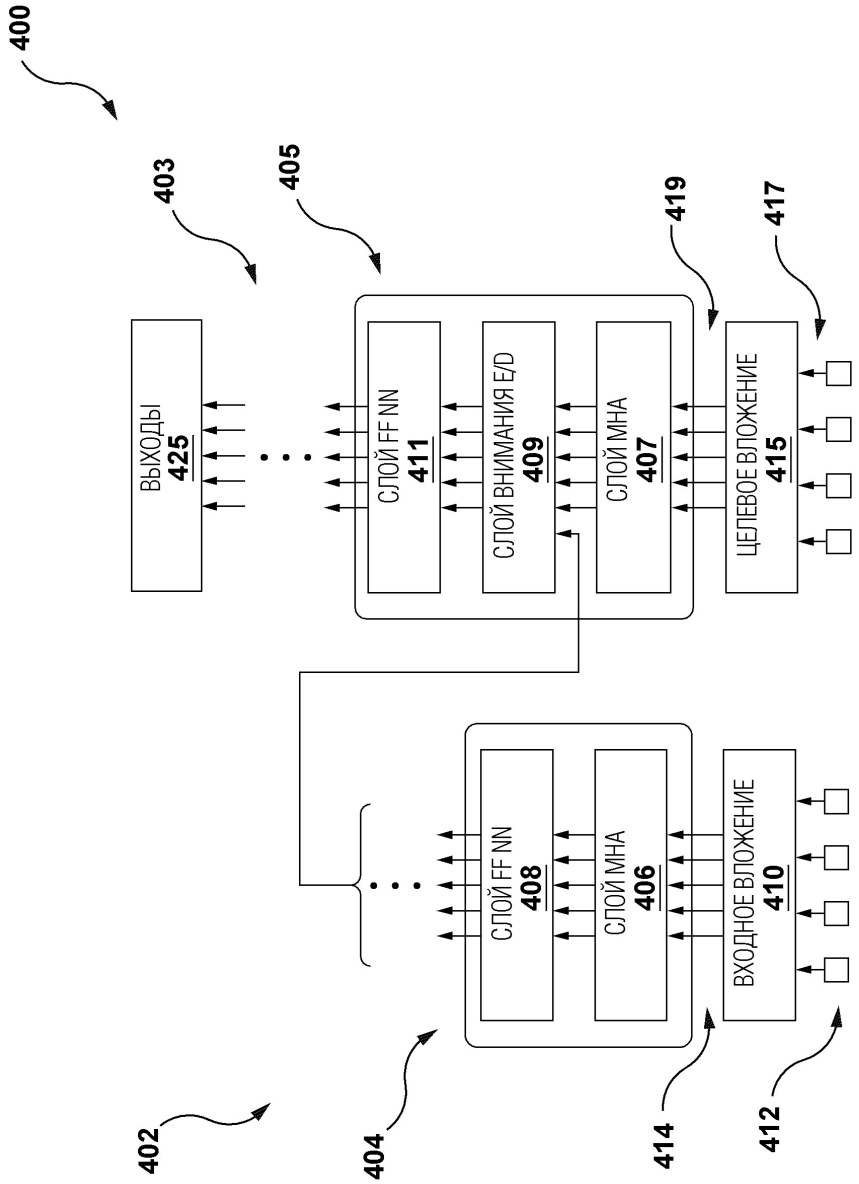
ФИГ. 2

3

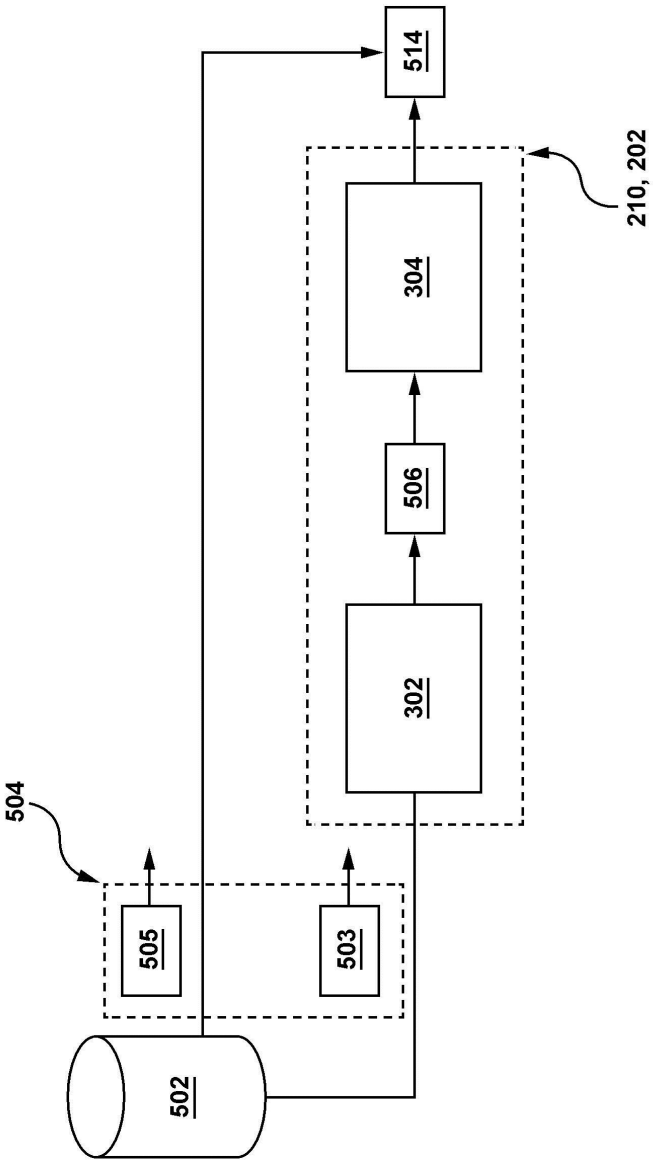


ФИГ. 3

4

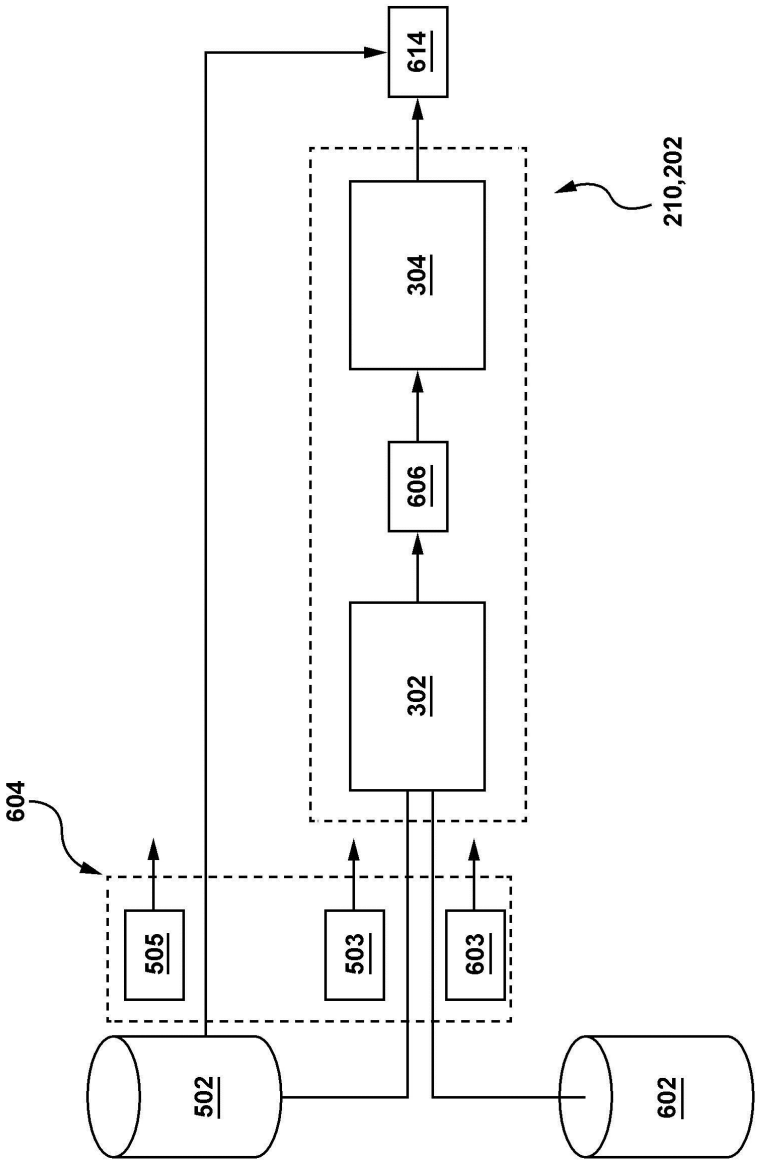

400
403
405
419
417
ВЫХОДЫ
425
СЛОЙ FF NN
411
СЛОЙ ВНИМАНИЯ E/D
409
СЛОЙ MHA
407
ЦЕЛЕВОЕ ВЛОЖЕНИЕ
415
СЛОЙ FF NN
408
СЛОЙ MHA
406
ВХОДНОЕ ВЛОЖЕНИЕ
410
402
404
414
412

ФИГ. 4



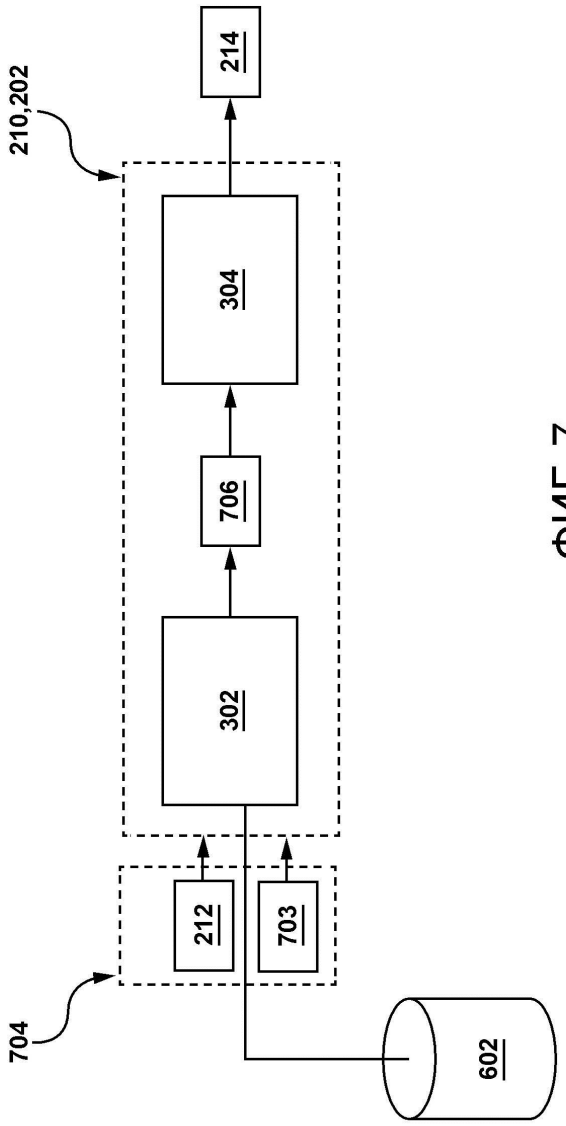
ФИГ. 5

6



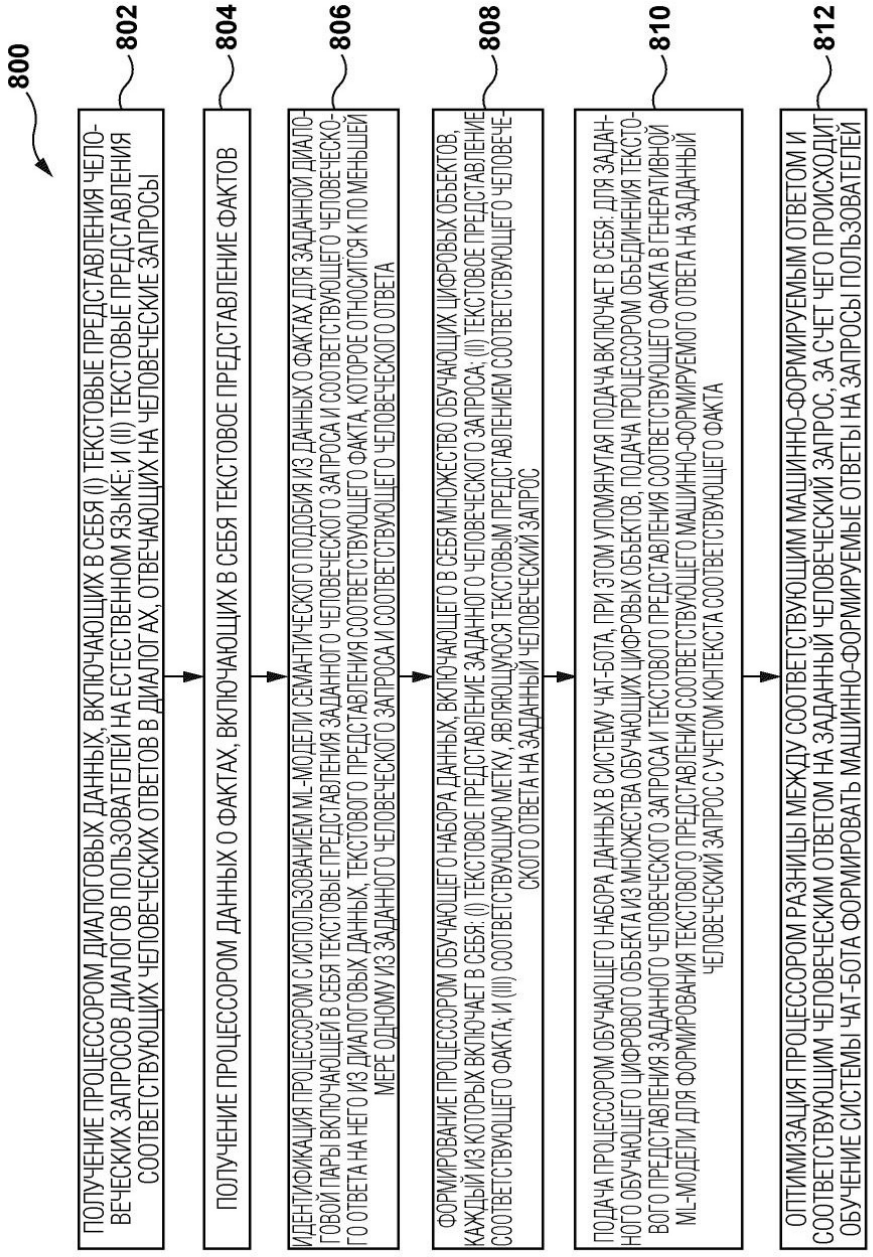
ФИГ. 6

7



ФИГ. 7

8


800
802
ПОЛУЧЕНИЕ ПРОЦЕССОРОМ ДИАЛОГОВЫХ ДАННЫХ, ВКЛЮЧАЮЩИХ В СЕБЯ (I) ТЕКСТОВЫЕ ПРЕДСТАВЛЕНИЯ ЧЕЛОВЕЧЕСКИХ ЗАПРОСОВ ДИАЛОГОВ ПОЛЬЗОВАТЕЛЕЙ НА ЕСТЕСТВЕННОМ ЯЗЫКЕ; И (II) ТЕКСТОВЫЕ ПРЕДСТАВЛЕНИЯ СООТВЕТСТВУЮЩИХ ЧЕЛОВЕЧЕСКИХ ОТВЕТОВ В ДИАЛОГАХ, ОТВЕЧАЮЩИХ НА ЧЕЛОВЕЧЕСКИЕ ЗАПРОСЫ
804
ПОЛУЧЕНИЕ ПРОЦЕССОРОМ ДАННЫХ О ФАКТАХ, ВКЛЮЧАЮЩИХ В СЕБЯ ТЕКСТОВОЕ ПРЕДСТАВЛЕНИЕ ФАКТОВ
806
ИДЕНТИФИКАЦИЯ ПРОЦЕССОРОМ С ИСПОЛЬЗОВАНИЕМ ML-МОДЕЛИ СЕМАНТИЧЕСКОГО ПОДОБИЯ ИЗ ДАННЫХ О ФАКТАХ ДЛЯ ЗАДАННОЙ ДИАЛОГОВОЙ ПАРЫ ВКЛЮЧАЮЩЕЙ В СЕБЯ ТЕКСТОВЫЕ ПРЕДСТАВЛЕНИЯ ЗАДАННОГО ЧЕЛОВЕЧЕСКОГО ЗАПРОСА И СООТВЕТСТВУЮЩЕГО ЧЕЛОВЕЧЕСКОГО ОТВЕТА НА НЕГО ИЗ ДИАЛОГОВЫХ ДАННЫХ, ТЕКСТОВОГО ПРЕДСТАВЛЕНИЯ СООТВЕТСТВУЮЩЕГО ФАКТА, КОТОРОЕ ОТНОСИТСЯ К ПО МЕНЬШЕЙ МЕРЕ ОДНОМУ ИЗ ЗАДАННОГО ЧЕЛОВЕЧЕСКОГО ЗАПРОСА И СООТВЕТСТВУЮЩЕГО ЧЕЛОВЕЧЕСКОГО ОТВЕТА
808
ФОРМИРОВАНИЕ ПРОЦЕССОРОМ ОБУЧАЮЩЕГО НАБОРА ДАННЫХ, ВКЛЮЧАЮЩЕГО В СЕБЯ МНОЖЕСТВО ОБУЧАЮЩИХ ЦИФРОВЫХ ОБЪЕКТОВ, КАЖДЫЙ ИЗ КОТОРЫХ ВКЛЮЧАЕТ В СЕБЯ: (I) ТЕКСТОВОЕ ПРЕДСТАВЛЕНИЕ ЗАДАННОГО ЧЕЛОВЕЧЕСКОГО ЗАПРОСА; (II) ТЕКСТОВОЕ ПРЕДСТАВЛЕНИЕ СООТВЕТСТВУЮЩЕГО ФАКТА; И (III) СООТВЕТСТВУЮЩУЮ МЕТКУ, ЯВЛЯЮЩУЮСЯ ТЕКСТОВЫМ ПРЕДСТАВЛЕНИЕМ СООТВЕТСТВУЮЩЕГО ЧЕЛОВЕЧЕСКОГО ОТВЕТА НА ЗАДАННЫЙ ЧЕЛОВЕЧЕСКИЙ ЗАПРОС
810
ПОДАЧА ПРОЦЕССОРОМ ОБУЧАЮЩЕГО НАБОРА ДАННЫХ В СИСТЕМУ ЧАТ-БОТА, ПРИ ЭТОМ УПОМЯНУТАЯ ПОДАЧА ВКЛЮЧАЕТ В СЕБЯ: ДЛЯ ЗАДАННОГО ОБУЧАЮЩЕГО ЦИФРОВОГО ОБЪЕКТА ИЗ МНОЖЕСТВА ОБУЧАЮЩИХ ЦИФРОВЫХ ОБЪЕКТОВ, ПОДАЧА ПРОЦЕССОРОМ ОБЪЕДИНЕНИЯ ТЕКСТОВОГО ПРЕДСТАВЛЕНИЯ ЗАДАННОГО ЧЕЛОВЕЧЕСКОГО ЗАПРОСА И ТЕКСТОВОГО ПРЕДСТАВЛЕНИЯ СООТВЕТСТВУЮЩЕГО ФАКТА В ГЕНЕРАТИВНОЙ ML-МОДЕЛИ ДЛЯ ФОРМИРОВАНИЯ ТЕКСТОВОГО ПРЕДСТАВЛЕНИЯ СООТВЕТСТВУЮЩЕГО МАШИННО-ФОРМИРУЕМОГО ОТВЕТА НА ЗАДАННЫЙ ЧЕЛОВЕЧЕСКИЙ ЗАПРОС С УЧЕТОМ КОНТЕКСТА СООТВЕТСТВУЮЩЕГО ФАКТА
812
ОПТИМИЗАЦИЯ ПРОЦЕССОРОМ РАЗНИЦЫ МЕЖДУ СООТВЕТСТВУЮЩИМ МАШИННО-ФОРМИРУЕМЫМ ОТВЕТОМ И СООТВЕТСТВУЮЩИМ ЧЕЛОВЕЧЕСКИМ ОТВЕТОМ НА ЗАДАННЫЙ ЧЕЛОВЕЧЕСКИЙ ЗАПРОС, ЗА СЧЕТ ЧЕГО ПРОИСХОДИТ ОБУЧЕНИЕ СИСТЕМЫ ЧАТ-БОТА ФОРМИРОВАТЬ МАШИННО-ФОРМИРУЕМЫЕ ОТВЕТЫ НА ЗАПРОСЫ ПОЛЬЗОВАТЕЛЕЙ

ФИГ. 8