(19) Евразийское патентное ведомство

(11) 047778 (13) B1

(12) ОПИСАНИЕ ИЗОБРЕТЕНИЯ К ЕВРАЗИЙСКОМУ ПАТЕНТУ

(45) Дата публикации и выдачи патента
2024.09.09

(21) Номер заявки
202393319

(22) Дата подачи заявки
2023.12.19

(51) Int. Cl. ***G06F 40/253*** (2020.01)
G06F 40/56 (2020.01)
G06F 40/40 (2020.01)
G06F 40/30 (2020.01)
G06F 40/10 (2020.01)

(54) СПОСОБ И СИСТЕМА ПЕРЕФРАЗИРОВАНИЯ ТЕКСТА

(31) **2023116467**
(32) **2023.06.22**
(33) **RU**
(43) **2024.09.05**

(71)(73) Заявитель и патентовладелец:
ПУБЛИЧНОЕ АКЦИОНЕРНОЕ ОБЩЕСТВО "СБЕРБАНК РОССИИ" (ПАО СБЕРБАНК) (RU)

(72) Изобретатель:
Феногенова Алена Сергеевна, Тихонова Мария Ивановна (RU)

(74) Представитель:
Герасин Б.В. (RU)

(56) US-B2-11314950
US-A1-2022121879
US-A1-2023137209
US-B2-10915712

(57) Настоящее изобретение в общем относится к области вычислительной техники, а в частности к способу и системе преобразования текста на основе генерации текста. Техническим результатом, достигающимся при реализации заявленного изобретения, является повышение оригинальности и семантической точности генерации перефразированного стилизованного текста из исходного текста. Указанный результат достигается за счет способа генерации текста в системе перефразирования текста, содержащего этапы, на которых: получают текстовый фрагмент на естественном языке; получают целевой стиль текстового фрагмента, характеризующий стилистические черты, присущие указанному целевому стилю, и параметр стилизации текста, характеризующий степень стилизации текста; обрабатывают текстовый фрагмент, причем обработка включает, по меньшей мере, разбиение указанного фрагмента на текстовые блоки; осуществляют кодирование каждого текстового блока, причем в ходе кодирования выполняют токенизацию текстового блока; выполняют векторизацию текстовых блоков по токенам; осуществляют обработку векторных представлений токенов каждого текстового блока, в ходе которой формируют набор кандидатов стилизованных перефразированных текстов для текстового блока в векторизированном виде; осуществляют декодирование каждого кандидата в каждом текстовом блоке, причем в ходе декодирования выполняют, по меньшей мере, преобразование векторизированных стилизованных текстов в токены и детокенизацию; осуществляют ранжирование набора кандидатов стилизованного перефразированного текста и выбор лучшего перефразированного стилизованного кандидата; объединяют стилизованные перефразированные тексты каждого блока с сохранением исходного порядка в перефразированный стилизованный текстовый фрагмент и отправляют указанный фрагмент в систему перефразирования текста.

047778 B1

047778 B1

Область техники

Заявленное изобретение в общем относится к области вычислительной техники, а в частности к способу и системе преобразования текста на основе генерации текста.

Уровень техники

В результате развития сферы информационных технологий, заметно увеличилось количество публикуемой текстовой информации. Так, в сети Интернет, расположено множество ресурсов с текстовыми данными, например, средства массовой информации, социальные сети, тематические ресурсы и т.д. Подготовкой и созданием таких текстовых данных занимаются специально обученные люди - копирайтеры. Так, в зависимости от места размещения целевого текста, исходные текстовые данные преобразуют в определенный стиль (деловой, научный, разговорный и т.д.) и определенный формат подачи, адаптируя и/или перефразируя текст с сохранением исходного смысла, обеспечивая при этом оригинальность полученного преобразованного текста. Например, такой процесс преобразования текста (перефразирования) особенно востребован в организациях, связанных с распространением информации, таких как СМИ (средство массовой информации), деятельность которых напрямую зависит от скорости и оригинальности созданного преобразованного текста на основе исходных текстовых данных. Кроме того, в крупных организациях, публикующих значительные объемы текста, также существует потребность в создании оригинального текста в определенном стиле. Так, автоматизация процесса преобразования текста или части процесса преобразования текста может позволить существенно повысить эффективность в таких сферах, как журналистика, редакторское дело, создание контента для медиа платформ и виртуальных ассистентов и т.д.

Однако, в настоящее время, создание такой технологии связано с рядом трудностей, не позволяющих осуществлять качественное перефразирование текста в заданном стиле с высокой точностью.

Так, из уровня техники известен ряд решений, направленных на генерацию преобразованного текста из исходного текста. Одним из подходов, применяющимся для преобразования текста, являются кодировщики - системы, которые изменяют процент плагиата и направлены на точечную замену и перестановку текса. Еще одним подходом, обеспечивающим преобразование текста, являются синонимайзеры - системы, которые меняют точечно синонимы из заготовленной заранее базы знаний.

К недостаткам таких решений можно отнести невозможность сохранения исходного смыслового содержания текста, потерю читабельности текста, примитивизмы и несогласованность текста, текстовую дегенерацию и непосредственно проблемы с перефразировкой на разных длинах текста. Кроме того, такого рода решения не предполагают и не предназначены для преобразования текста в заданном стиле изложения текста.

Из уровня техники также известно решение, раскрытое в заявке на патент США № US 2017/0132208 (IBM), опубл. 11.05.2017. Указанное решение раскрывает способ перефразирования текста, обеспечивающий возможность перефразирования текста в персональный стиль на основе заранее собранного словаря с терминами. Так, указанное решение выполнено с возможностью поиска и сопоставления схожих по смыслу слов и/или фраз из текста и собранного словаря и перефразирования получаемого текста путем замены найденных слов на слова или фразы из словаря.

Недостатком указанного решения является низкая степень преобразования (перефразирования) текста из-за применения только точечных замен и перестановок в тексе, что, как следствие, приводит к низкой точности преобразования текста и низкой оригинальности перефразированного текста. Также, указанное решение обеспечивает низкую точность стилизации текста при его перефразировании, так как предназначено только для стилизации, основанной на словаре, что, как следствие, не позволяет стилизовать перефразированный текст в разные стили. Кроме того, указанное решение не предполагает и не предназначено для перефразирования текстов большой длины, так как такой процесс требует обеспечение возможности перефразирования текста на уровне всего документа, т.е. учета связей между предложениями.

Общими недостатками существующих решений является отсутствие эффективного способа преобразования текста с сохранением исходного смысла в заданном стиле с высокой точностью. Кроме того, такого рода решение должно обеспечивать высокую степень оригинальности перефразированного текста в заданном стиле и обеспечивать перефразирование текстов большой длины с учетом связей между предложениями. Также, такого рода решение должно обеспечивать автоматическую генерацию текста разного характера в заранее заданном стиле с той или иной степенью выраженности, сохраняя смысл при значимом проценте оригинальности полученного текста.

Раскрытие изобретения

В заявленном изобретении предлагается новый подход к генерации текста для систем преобразования текста. В данном решении используются алгоритмы машинного обучения, которые позволяют осуществлять генерацию стилизованного преобразованного текста из исходного текста с высокой точностью, обеспечивающей семантическую близость исходного и перефразированного текста, исключают искажение текста новыми фактами и обеспечивают высокую степень оригинальности перефразированного текста.

Таким образом, решается техническая проблема обеспечения возможности генерации перефразиро-

ванного стилизованного текста с высокой точность.

Техническим результатом, достигающимся при решении данной проблемы, является повышение оригинальности и семантической точности генерации перефразированного стилизованного текста из исходного текста.

Дополнительным техническим результатом, проявляющимся при решении вышеуказанной проблемы, является обеспечение возможности перефразирования текстов в заданном стиле с сохранением исходного смыслового содержания.

Указанные технические результаты достигаются благодаря осуществлению способа генерации текста в системе перефразирования текста, выполняемого по меньшей мере одним вычислительным устройством, и содержащего этапы, на которых:

a) получают текстовый фрагмент на естественном языке;

b) получают целевой стиль текстового фрагмента, характеризующий стилистические черты, присущие указанному целевому стилю, и параметр стилизации текста, характеризующий степень стилизации текста;

c) обрабатывают текстовый фрагмент, причем обработка включает по меньшей мере разбиение указанного фрагмента на текстовые блоки;

d) осуществляют кодирование каждого текстового блока, причем в ходе кодирования выполняют токенизацию текстового блока;

e) выполняют векторизацию текстовых блоков по токенам, полученных на этапе d);

f) осуществляют обработку векторных представлений токенов каждого текстового блока, полученных на этапе e), в ходе которой:

i) предсказывают вероятность следующего токена для каждого векторизованного текстового представления токенов текстового блока с помощью первой модели машинного обучения на базе нейронной сети;

ii) обрабатывают данные, полученные на шаге i, с помощью второй модели машинного обучения на базе нейронной сети, в ходе которой осуществляют перевзвешивание и корректировку предсказания первой модели, в ходе которой перевзвешивают вероятности следующего токена, с учетом параметра стилизации текста;

iii) генерируют следующий токен для каждого векторизованного текстового представления токенов текстового блока на основе распределения вероятностей следующего токена, полученных на шаге ii, и добавляют указанный токен в конец векторизованного текстового представления токенов блока;

iv) генерируют векторизованное текстовое представление стилизованного перефразированного текста для текстового блока, итеративно повторяя шаги i-iii до первого критерия останова;

v) генерируют по меньшей мере один стилизованный перефразированный текст для текстового блока в векторизированном виде итеративно повторяя шаги i-iv до второго критерия останова;

vi) формируют набор кандидатов стилизованных перефразированных текстов для текстового блока в векторизированном виде на основе данных, полученных на шаге v;

g) осуществляют декодирование каждого кандидата, полученного на этапе f), в каждом текстовом блоке, причем в ходе декодирования выполняют по меньшей мере преобразование векторизированных стилизованных текстов в токены и детокенизацию;

h) осуществляют ранжирование набора кандидатов стилизованного перефразированного текста и выбор лучшего перефразированного стилизованного кандидата, причем выбор лучшего перефразированного стилизованного кандидата основан на попарном расстоянии между исходным текстовым фрагментом блока и каждым из возможных перефразированных стилизованных текстовых фрагментов из набора;

i) объединяют стилизованные перефразированные тексты каждого блока с сохранением исходного порядка в перефразированный стилизованный текстовый фрагмент и отправляют указанный фрагмент в систему перефразирования текста.

В одном из частных вариантов реализации способа текстовый блок не превышает заранее заданный параметр длины блока.

В другом частном варианте реализации способа текстовый фрагмент разбивается на текстовые блоки не превышающие заданный параметр длины блока.

В другом частном варианте реализации способа первый критерий останова представляет собой максимальную допустимую длину текстового блока или символ, соответствующий концу предложения.

В другом частном варианте реализации способа допустимая длина текстового блока не превышает 200 слов.

В другом частном варианте реализации способа второй критерий останова представляет собой целочисленное значение, характеризующее требуемое количество кандидатов.

В другом частном варианте реализации способа текстовый фрагмент содержит по меньшей мере два предложения.

В другом частном варианте реализации способа текстовые блоки состоят из законченного числа предложений.

Кроме того, заявленные технические результаты достигаются за счет системы генерации текста, со-

держащей:

по меньшей мере один процессор;

по меньшей мере одну память, соединенную с процессором, которая содержит машиночитаемые инструкции, которые при их выполнении по меньшей мере одним процессором обеспечивают выполнение способа генерации текста.

Кроме того, заявленные технические результаты достигаются за счет способа перефразирования текста, выполняемого по меньшей мере одним вычислительным устройством, и содержащего этапы, на которых:

a) получают запрос пользователя на перефразирование текстового фрагмента на естественном языке и целевой стиль текста;

b) обрабатывают, полученный на этапе a), текстовый фрагмент, с помощью способа генерации текста в системе перефразирования текста;

c) отображают ответ на запрос пользователя, содержащий перефразированный стилизованный текстовый фрагмент, полученный на этапе b).

В одном из частных вариантов реализации способа текстовый фрагмент на естественном языке содержит по меньшей мере два предложения.

Кроме того, заявленные технические результаты достигаются за счет системы перефразирования текста, содержащей:

по меньшей мере один процессор;

по меньшей мере одну память, соединенную с процессором, которая содержит машиночитаемые инструкции, которые при их выполнении по меньшей мере одним процессором обеспечивают выполнение способа перефразирования текста.

Краткое описание чертежей

Признаки и преимущества настоящего изобретения станут очевидными из приводимого ниже подробного описания изобретения и прилагаемых чертежей.

Фиг. 1 иллюстрирует систему перефразирования текста.

Фиг. 2 иллюстрирует блок-схему выполнения заявленного способа.

Фиг. 3 иллюстрирует блок-схему алгоритма генерации перефразированных стилизованных текстов.

Фиг. 4 иллюстрирует пример общего вида вычислительного устройства, которое обеспечивает реализацию заявленного решения.

Осуществление изобретения

Ниже будут описаны понятия и термины, необходимые для понимания данного технического решения.

Модель в машинном обучении (МО) - совокупность методов искусственного интеллекта, характерной чертой которых является не прямое решение задачи, а обучение в процессе применения решений множества сходных задач.

Распознавание именованных сущностей (Named-entity recognition, NER) - это подзадача извлечения информации, которая направлена на поиск и классификацию упоминаний именованных сущностей в неструктурированном тексте по заранее определенным категориям, таким как имена собственные, имена персонажей, организации, местоположения, денежные значения, проценты и т.д.

Векторное представление слов (word embeddings, эмбеддинги) - общее название для различных подходов к моделированию языка и обучению представлений в обработке естественного языка, направленных на сопоставление словам (и, возможно, фразам) из некоторого словаря векторов из n-мерного вещественного пространства R_n.

Токенизация - это процесс разбиения текста на текстовые единицы или токены (чаще всего в качестве таких единиц выступают слова, но это могут быть также буквы, части предложения, сочетания слов и т.д.).

Языковая модель - это вероятностное распределение на множестве словарных последовательностей. В данном патенте термин "языковая модель" употребляется для описания нейросетевых языковых моделей, которые выполнены с возможностью моделирования языка посредством оценки вероятности той или иной последовательности символов.

Авторегрессионная (языковая) модель - в общем случае это модель временного ряда, в которой предсказания для временного ряда зависят от предыдущих предсказаний временного ряда. В применении к нейросетевым языковым моделям этот термин употребляется в значении того, что предсказания для следующего токена делается на основе предыдущих.

Парафраз - преобразование текста с сохранением исходного смысла. Под парафразом текста стоит понимать изложение текста другими словами с сохранением смыслового содержания исходного текста. Стоит отметить, что в данном патенте термин парафраз, парафразирование и перефразирование являются эквивалентными и описывают один и тот же процесс, и являются взаимозаменяемыми.

Под стилизацией текста в данном решении понимается генерирование текста путем преобразования принятого исходного текста в текст, которому присущи стилистические речевые черты. Так, стилистическими чертами могут являться эмоциональный окрас текста (веселый, грустный и т.д.). В другом частном

варианте осуществления стилистическими речевыми чертами может являться условие и цели общения в какой-то сфере общественной деятельности, например, официально-деловой деятельности, публицистической деятельности, разговорной, художественной и т.д. Стоит отметить, что стилизацией текста также может являться придание характерных черт тексту, присущих особенностям общения отдельно взятых личностей, персонажей, литературных героев и т.д., не ограничиваясь. Так, в еще одном частном варианте осуществления стилизацией текста может являться преобразование исходного текста в стилизованный, в соответствии с заданным стилем общения конкретного цифрового помощника, которому присущи разговорные черты стиля, например, использование определенных местоимений, подчеркивающих неформальный/формальный стиль общения ("Ты" и "Вы"), род, число и т.д.

Заявленное изобретение предлагает новый подход, обеспечивающий повышение оригинальности перефразированного текста в заданном стиле из исходного текста при сохранении исходного смыслового содержания. Так, особенности заявленного технического решения, обеспечивают возможность получения текстов разного характера в заранее заданном стиле с той или иной степенью выраженности, сохраняя при этом семантическую точность при значимом проценте оригинальности полученного текста. Оригинальность - изменение структуры текста без изменения смыслового содержания. Кроме того, указанное техническое решение обеспечивает возможность перефразирования текстов большой длины с учетом связей между предложениями. Также, еще одним преимуществом, достигающимся при использовании заявленного изобретения, является возможность стилизации перефразированного текста в любое число стилей, за счет внедрения модели машинного обучения, обученной на корректировку перефразированного текста в зависимости от заданного параметра стиля. Также, возможность перефразирования текста в заданном стиле обеспечивает автоматизацию процесса генерации перефразированного стилизованного текста, что значительно снижает время, требуемое на перефразирование текста и дальнейшую стилизацию по сравнению с ручной генерацией каждого отдельного сегмента (перефразирования и стилизации).

Заявленное техническое решение может быть реализовано на компьютере, в виде автоматизированной информационной системы (АИС) или машиночитаемого носителя, содержащего инструкции для выполнения вышеупомянутого способа.

Техническое решение также может быть реализовано в виде распределенной компьютерной системы или вычислительного устройства.

В данном решении под системой подразумевается компьютерная система, ЭВМ (электронно-вычислительная машина), ЧПУ (числовое программное управление), ПЛК (программируемый логический контроллер), компьютеризированные системы управления и любые другие устройства, способные выполнять заданную, чётко определённую последовательность вычислительных операций (действий, инструкций).

Под устройством обработки команд подразумевается электронный блок либо интегральная схема (микропроцессор), исполняющая машинные инструкции (программы).

Устройство обработки команд считывает и выполняет машинные инструкции (программы) с одного или более устройства хранения данных, например, таких устройств, как оперативно запоминающие устройства (ОЗУ) и/или постоянные запоминающие устройства (ПЗУ). В качестве ПЗУ могут выступать, но, не ограничиваясь, жесткие диски (HDD), флэш-память, твердотельные накопители (SSD), оптические носители данных (CD, DVD, BD, MD и т.п.) и др.

Программа - последовательность инструкций, предназначенных для исполнения устройством управления вычислительной машины или устройством обработки команд.

Термин "инструкции", используемый в этой заявке, может относиться, в общем, к программным инструкциям или программным командам, которые написаны на заданном языке программирования для осуществления конкретной функции, такой как, например, кодирование и декодирование текстов, фильтрация, ранжирование, трансляция текстов в диалоговую систему и т.п. Инструкции могут быть осуществлены множеством способов, включающих в себя, например, объектно-ориентированные методы. Например, инструкции могут быть реализованы, посредством языка программирования Python, C++, Java, Python, различных библиотек (например, MFC; Microsoft Foundation Classes) и т.д. Инструкции, осуществляющие процессы, описанные в этом решении, могут передаваться как по проводным, так и по беспроводным каналам передачи данных, например Wi-Fi, Bluetooth, USB, WLAN, LAN и т.п.

На фиг. 1 приведен общий вид системы 100 перефразирования текста. Система 100 включает в себя основные функциональные элементы, такие как: модуль предобработки текста 101, модуль кодирования/декодирования 102, модуль стилевого перефразирования текста 103, модуль ранжирования 104 перефразированных стилизованных текстов, модуль постобработки текста 105. Более подробно элементы системы 100 раскрыты на фиг. 4.

Система перефразирования текста 100 может использоваться в таких сферах обработки текста, как: редактирование, подготовка новостных статей, презентаций, подготовка текста для цифровых ассистентов, диалоговых систем и т.д. Система 100 может быть интегрирована в процессы создания текстового контента организациями и обеспечивать высокую скорость генерации перефразированных стилизованных текстов при сохранении высокой степени оригинальности текста. Кроме того, использование системы 100 может обеспечивать изложение текстовых данных в едином уникальном стиле. Такая функция

востребована в больших организациях, где имеется штат рерайтеров, выполняющих аналогичные функции, но при этом сталкивающихся с проблемами формирования текстов с учетом человеческого фактора и стиля, что, как следствие, требует дополнительной обработки текстов для корректировки их в единый стиль. Соответственно, система 100 способна решить указанную проблему. Кроме того, система 100 может являться частью диалоговой системы и/или может быть связана с диалоговой системой посредством каналов связи, таких как Интернет. При общении пользователя с диалоговой системой, указанная диалоговая система может обращаться к системе 100 с целью обработки пользовательского запроса и получения текстовых данных, которые будут отображены пользователю. В еще одном частном варианте осуществления, система 100 может представлять собой сервер, оборудованный модулями, обеспечивающими выполнение действий, направленных на получение перефразированного стилизованного текста из исходного текста. Система 100 предназначена для генерации из исходных текстов различной длины, в том числе и большой длины (десятки страниц), перефразированные стилизованные тексты с высоким процентом оригинальности. Стоит отметить, что под генерацией перефразированного стилизованного текста понимается генерация семантически идентичного, по отношению к исходному, текста в котором изменяется конструкция и формулировка слов (в том числе с применением синонимов) в определенном речевом стиле. Т.е. особенностью настоящего технического решения является возможность перефразирования исходного текста и придания такому тексту стилевых черт заданного стиля.

Под пользователем 110 в данном решении следует понимать как и пользователя системы 100, например, пользователя персонального вычислительного устройства, такого как компьютер, ноутбук, способного отправлять исходный текст и получать результаты работы системы 100, так и автоматизированные средства приема/получения входных/выходных данных (результатов генерации текста) от системы 100. Так, например, пользователем 110 может являться удаленный сервер организации, соединенный с системой 100 по каналам передачи данных.

Модуль предобработки 101 может быть реализован на базе по меньшей мере одного вычислительного устройства, оснащенного соответствующим программным обеспечением, и используется для предобработки входного текста, полученного, например, от пользователя, а также для разбиения длинных текстов на отдельные текстовые фрагменты, которые затем попадаются в следующий модуль кодирования/декодирования. Хотя в настоящем техническом решении упоминается получения текстовых данных от пользователя, следует понимать, что под пользователем понимается любой канал взаимодействия с системой 100. Так, получение текстовых данных может выполняться посредством машинного взаимодействия внешней автоматизированной системы общения с пользователем с системой 100, и т.д.

Более подробно, модуль 101, выполнен с возможностью получения на вход исходных текстовых данных, как в формате запроса пользователя, так и в формате текстового файла, от пользователя, обработки текстовых данных для удаления из них спецсимволов, лишних пробелов, лишних повторов, и разбиения обработанного текста на отдельные текстовые фрагменты.

Обработка текстовых данных может выполняться, например, посредством регулярных выражений. Кроме того, также могут применяться автоматические методы, основанные на регулярных выражениях и программно-аппаратные средства проверки пунктуации, например, Microsoft Office®, Google® и т.д.

Разбиение длинных текстов на отдельные текстовые фрагменты, может происходить на основе конца предложения и/или максимальной допустимой длины текстового блока, например, 200 слов. При разбиении исходный текст разделяется на отдельные текстовые блоки. Разбиение происходит по предложениям так, чтобы в каждом блоке был законченный набор предложений (т.е. предложение посередине не разбивается, чтобы не получилось так, что половина попала в один фрагмент, а половина в другой). В случае, если произвести разбиение так, чтобы в каждом текстовом фрагменте оказался законченный набор предложений невозможно, то разбиение происходит по максимально допустимой длине блока.

В еще одном варианте осуществления, для ускорения работы системы допускается обработка текстовых фрагментов по батчам, то есть наборами (или пакетами) в определенном количестве.

Модуль кодирования/декодирования 102 может быть реализован на базе по меньшей мере одного вычислительного устройства, оснащенного соответствующим программным обеспечением включать набор моделей для токенизации и детокенизации текста, векторизации токенизированного текста и преобразования токенов в текст, например, одну или несколько моделей машинного обучения для преобразования текстовой информации в векторную форму, например, BERT, ELMo, ULMFit, XLNet, RoBerta, RuGPT3 и другие. Модуль кодирования/декодирования выполнен с возможностью приема на вход предобработанных текстовых блоков от модуля 101. В одном частном варианте осуществления модуль 101 может быть реализован на базе системы 400, которая более подробно раскрыта на фиг. 4. Стоит отметить, что определенный метод токенизации и векторизации зависит от выбранной языковой модели, на базе которой реализован модуль 102. Например, при использовании модели RuGPT3, токенизация осуществляется методом BPE (Byte Pair Encoding), а последующая векторизация - путем замены каждого токена на его индекс в словаре языковой модели, составленном на этапе изначального обучения модели. Кроме того, в еще одном частном варианте осуществления, в качестве метода токенизации может использоваться токенизация по словам. Пример токенизации по словам и кодирование слов индексами в словаре:

'мама мыла раму' → ['мама', 'мыла', 'раму'] → [235, 376, 1056]

Модуль стилевого перефразирования текста 103 может быть реализован на базе по меньшей мере двух нейронных сетей, заранее обученных на конкретных наборах данных, соответствующих решаемой подзадаче (перефразирование и стилизация). Модуль 103 предназначен для преобразования (перефразирования) текстового фрагмента в заданном стиле. Модуль принимает на вход закодированный модулем 102 текстовый фрагмент (в общем виде набор текстовых блоков), стиль, в котором будет перефразирован текст и коэффициент силы стиля (коэффициент выраженности стиля) $w \geq 0$ (чем больше w, тем сильнее выражен стиль). В одном частном варианте осуществления, модуль 103 состоит из двух подмодулей: подмодуль перефразирования, стилевой подмодуль. При генерации парафразированного текста эти два подмодуля взаимодействуют между собой в виде так называемой управляемой текстовой генерации. Стилевой подмодуль управляет генераций подмодуля перефразирования, добиваясь того, чтобы генерация происходила в заданном стиле.

Рассмотрим более подробно указанные подмодули.

Подмодуль перефразирования может быть реализован на базе по меньшей мере одной нейронной сети, заранее обученной на конкретных наборах текстов. Для дообучения могут использоваться пары "исходный текст" - "перефразированный текст". В качестве модели машинного обучения, реализующей функцию генерации перефразированных текстов, может быть использована, например, генеративная языковая модель, такая как RuT5, RuGPT3, XLNet и т.д. В одном частном варианте осуществления, при реализации заявленного решения, МО являлась русскоязычная генеративная языковая модель RuT5-Large https://huggingface.co/sberbank-ai/ruT5-large. Модель обучена на источниках из разных доменов: Википедия, книги, новости, русский Common Crawl и т.д. На данном этапе обучения, результатом обучения языковой модели являлась возможность предсказывания вероятности следующего токена на основе предыдущего начального фрагмента текста: $p(y_t|y_{1:t-1})$, где y_t - предсказываемый следующий токен, $y_{1:t-1}$ - начальный фрагмент текста.

Так, если в процессе обучения модель часто встречала в обучающих данных определенное словосочетание, то при предсказании следующего после известного из словосочетания токена, модель с высокой вероятностью будет предсказывать именно токен из словосочетания в обучающем наборе данных.

Далее, для выполнения непосредственно самого процесса генерации перефразированного текста из исходного текстового фрагмента, выполнялось дообучение обученной модели. Для дообучения модели использовалась процедура fine-tune. На указанном этапе выполнялось настраивание весов обученной модели в соответствии с решаемой задачей. Так, при появлении в модели исходного текста, за счет измененных весовых коэффициентов, наиболее вероятным предложением для модели будет рерайт (= парафраз) исходного текста (= исходный текст парафразированный другими словами), т.е. осуществляется повышение вероятности продолжения текстового фрагмента в том формате, в котором должен быть текст. Это можно описать в следующем виде. В выражении $p(y_t|y_{1:t-1},x)$ (где x - исходный парафразируемый текстовый фрагмент, y_t -предсказываемый следующий токен, $y_{1:t-1}$ - начальный префикс парафразированного текста, сгенерированный моделью) повышается вероятность токенов y_t так, чтобы сгенерированный префикс текста $y_{1:t-1}$ был начальным префиксом парафраза исходного текста x. Стоит отметить, что парафраз - это изменение структуры текста с сохранением смысла, в том числе с применением синонимов. Т.е. дообучение МО заключалось в обеспечении возможности генерации перефразированного текста за счет возможности перефразирования (выражения текста другими словами), в том числе с применением синонимов.

Так, при дообучении модели использовались датасеты из разных доменов.

Более подробно, дообучение выполнялось на текстах как разной длины, так и из разных доменов, для обеспечения возможности применения в разнообразных целевых задачах, например, контент сайтов, новости, отзывы, диалоги и многое другое.

Обучающий набор включал в себя готовые произведения разных авторов, переведённые с одних языков на другие, тексты, размеченные вручную в разных жанрах, а именно художественная литература, новости, отзывы, комментарии, разговорные реплики. Кроме того, указанный набор также содержал автоматически переведенные тексты, причем, отбирались только те тексты, которые соответствовали критерию фильтрации (использовались метрики качества Bertscore и Rouge-L).

Так, для обучающего набора, в частности, использовались предложения из следующих источников tapaco (русскоязычная часть, отфильтрованная по длинам), ParaphraserPlus (отфильтрованные предложения по Bertscore и длине), предложения разных авторов, переведённые с одних языков на другие языки.

Также, стоит отметить, что в процессе дообучения модели, применялся подход, основанный на обучении пар (оригинальный текст → целевой текст) в обе стороны, что обеспечило существенное снижение ошибок модели в части изменения длины перефразированного текста. При этом, такой подход также обеспечил увеличение количества примеров (аугментация набора данных).

Всего, в ходе дообучения модели, было использовано около 7000 примеров текстов разной длины и разных доменов.

Таким образом, итоговая дообученная модель имеет следующие метрики качества. Для пар [ориги-

нальный текст; сгенерированный текст] были использованы следующие автоматические метрики (см. табл. 1): 1) Mean Bleu - средняя оценка по всем текстам метрики BLEU (BLEU-1); 2) Mean Rouge - средняя оценка по всем текстам метрики ROUGE (ROUGE-L); 3) Bert score - средняя оценка по всем текстам метрики BertScore; 4) Mean labse score - средняя оценка по всем текстам метрики LABSE; 5) Sentence repeat - процент предложений схожих с оригинальным текстом. Более подробно указанные метрики раскрыты ниже.

Таблица 1

	BERTscore	Bleu	LABSE	Rouge – L	Sentence repeat
Подмодуль перефразирования	0.77	0.15	0.852	0.42	0.019

Как видно из табл. 1, высокие показатели метрик LabSE и BertScore говорят о том, что полученное предложение по смыслу близко к исходному.

Также помимо автоматических метрик была произведена оценка сгенерированных парафразов живыми людьми (кандидаты выбрались случайным образом и оценивали качество параграфов с помощью опросов) (табл. 2). Кандидаты оценивали насколько сгенерированный текст 1) грамматичен (Grammar); 2) оригинален (Originality); 3) передает смысл (Meaning). Результаты оценки.

Таблица 2

	Grammar	Meaning	Originality
подмодуль перерфразирования	0.92	0.74	0.87

Как видно из табл. 2, такие высокие значения результатов человеческой оценки говорят о высоком качестве парафразов, сгенерированных моделью: большой процент грамматической корректности (grammar) и высокая оригинальность полученных текстов (originality).

Далее рассмотрим стилевой подмодуль.

Указанный стилевой подмодуль может быть реализован на базе по меньшей мере одной нейронной сети, заранее дообученной под задачу классификации заданного набора стилей. Такой классификатор стилей затем используется для управления генерацией подмодуля перефразирования, чтобы направлять генерацию парафраза в русло заданного стиля.

В качестве модели машинного обучения, реализующей "управление" моделью перефразирования, в соответствии с заданным стилем, текстов, может быть использована, например, генеративная языковая модель, такая как RuT5, RuGPT3, XLNet и т.д. В одном частном варианте осуществления, при реализации заявленного решения, МО являлась русскоязычная генеративная языковая модель RuT5-Large https://huggingface.co/sberbank-ai/ruT5-large (та же, что и для подмодуля перефразирования).

Модель может поддерживать разнообразный набор стилей в зависимости от данных, на которых она была обучена. Например, позитивный, негативный, разговорный, официальный, грубый, книжный и другие. В данном случае для решения задачи классификации стиля выполнялось дообучение обученной модели под задачу классификации стиля по текстовому фрагменту. В одном частном случае выполнялось обучение модели под задачу бинарной классификации: позитивный стиль, негативный. Пример задачи: задача бинарной классификации по тексту предсказать является ли он позитивным или негативным. "Какой ужасный день"- Label: негатив. "Какой великолепный день" - Label: позитив. Для дообучения модели использовалась процедура fine-tune. На указанном этапе выполнялось настраивание весов обученной модели в соответствии с решаемой задачей. Так, после подобного дообучения модель, за счет измененных весовых коэффициентов, для заданного текстового фрагмента способна предсказывать его стиль (например, негатив/позитив). $p_{st}(style|y_{1:t})$, где style - стиль текста, $y_{1:t}$ - префикс (то есть начало) текста.

Текущий подход к дообучению был основан на стандартном fine-tune подходе https://github.com/patil-suraj/exploring-T5/blob/master/t5_fine_tuning.ipynb для модели данной архитектуры (а именно T5), однако в данном случае при fine-tune использовалась особая функция штрафа. Более подробно, указанная функция штрафа описана в статье https://arxiv.org/abs/2109.08914. Она состоит из двух компонент: стандартного "генеративного лосса", как в классической процедуре fine-tune, и дополнительного "дискриминативного лосса", который дополнительно подталкивает модель разделять вероятности разных стилей между собой:

$$\mathcal{L}_G = -\frac{1}{N}\sum_{i=1}^{N}\frac{1}{T_i}\sum_{t=1}^{T_i}\log P(y_t^{(i)}|y_{<t}^{(i)}, c^{(i)})$$

$$\mathcal{L}_D = -\frac{1}{N}\sum_{i=1}^{N}\log P(c^{(i)}|y_{1:T_i}^{(i)})$$

$$\mathcal{L}_{ParaGeDi} = \lambda\mathcal{L}_D + (1-\lambda)\mathcal{L}_G$$

Для дообучения данной модели использовались следующие наборы данных из открытых источников: Rusentitweet Rusentiment Корпус Казакских новостей, RuReviews.

Все вышеперечисленные датасеты представляют собой наборы пар ("текст" - "лейбл").

"Какой ужасный день"- Label: негатив.

"Какой великолепный день" - Label: позитив.

Таким образом, итоговый датасет для дообучения модели представлял собой набор данных для задачи бинарной классификации (например, позитивный/негативный стиль текста, разговорный/не разговорный стиль текста, научный/не научный и т.д.). Взаимодействие подмодулей модуля 103 выполнялось следующим образом. Так, при генерации текста, подмодуль стилизации управляет генерацией подмодуля перефразирования текста следующим образом: на каждом шаге подмодуль перефразирования предсказывает вероятность следующего токена, а подмодуль стилизации осуществляет перевзвешивание предсказанных вероятностей следующего токена, с учетом заданного стиля и, например, коэффициента выраженности стиля, придавая большую вероятность токенам (и как следствие словам) заданного целевого стиля. Более подробно процесс работы модуля 103 раскрыт ниже. Таким образом, модуль 103, посредством взаимодействия подмодулей перефразирования и стилизации, обеспечивает возможность генерации перефразированного в заданном речевом стиле текста.

Оценка качества итоговых двух моделей машинного обучения, обеспечивающих стилизованное перефразирование текста производилась на текстовом сете, содержащем 100 предложений по разным тематикам. Оценка производилась по следующим метрикам: 1) BLEU (Papineni, K.; Roukos, S.; Ward, T.; Zhu, W. J. (2002). BLEU: a method for automatic evaluation of machine translation (PDF). ACL-2002: 40th Annual meeting of the Association for Computational Linguistics, pp. 311-318. CiteSeerX 10.1.1.19.9416.). Алгоритм BLEU сравнивает фразы двойного перевода с фразами, которые он находит в эталонном варианте, и взвешенно подсчитывает количество совпадений. Эти совпадения не зависят от позиции. Высшая степень совпадения указывает на более высокую степень сходства с эталонным переводом и более высокий балл. Внятность и грамматика не учитываются. 2) LaBSE - семантическая метрика. Данная метрика оценивает косинусное сходство между векторными представлениями предложений, полученными с помощью модели, которое соответствует семантической близости, более подробно данная метрика раскрыта по ссылке: https://huggingface.co/setu4993/LaBSE. 3) Rouge (Recall-Oriented Understudy for Gisting Evaluation) - представляет собой набор показателей, используемый для оценки автоматического суммирования и машинного перевода при обработке естественного языка. Метрики сравнивают автоматически созданную сводку или перевод со справочной информацией или набором ссылок (созданным человеком) сводкой или переводом. 4) BertScore - данная метрика вычисляет семантическую близость двух предложений, суммируя косинусную близость между эмбеддингами их токенов. Далее вычисляется F1 мера, (см., например, https://arxiv.org/abs/1904.09675).

Результаты оценки модуля стилизованного перефразирования, приведены в табл. 3 и 4.

Оценка перефразированного в позитивном стиле текста.

Таблица 3

Коэффициент выраженности стиля	Rouge	BLEU	Labse	BertScore
0	0.674578	47.461442	0.907119	0.898835
10	0.550047	32.488431	0.870006	0.860456
15	0.407762	20.058734	0.798158	0.813747
20	0.303897	12.255100	0.719628	0.767543

Оценка перефразированного в отрицательном стиле текста.

Таблица 4

Коэффициент выраженности стиля	Rouge	BLEU	Labse	BertScore
0	0.674065	46.932319	0.911766	0.901344
10	0.550047	32.713438	0.868233	0.868656
15	0.397318	18.169413	0.786843	0.811531
20	0.261834	9.251190	0.709408	0.766116

Из табл. 3 и таблицы 4 видно, что для коэффициента 0 - bleu около 50, для коэффициента 10, 15 - bleu около 40-45, для коэффициента 20 - bleu около 25, для коэффициента 10, 15 - Labse, BertScore около 85-90, для коэффициента 20 - Labse, BertScore около 75-80.

Такие значения BLEU для коэффициентов стиля 10 и 15 свидетельствуют о высокой оригинальности полученного текста, при этом метрики LabSE и BertScore говорят о том, что полученное предложение по смыслу близко к исходному.

При этом можно сказать, что для коэффициента 20, выраженность стиля становится слишком сильной, что может негативно повлиять на сохранение смысла. Также, в одном частном варианте осуществления использовался дополнительный анализ стилизации с помощью автоматической детекции выраженности позитива/негатива с помощью модели из библиотеки Достоевский (колонка mean_neg_score - Negativity, mean_pos_score - Positivity). Более подробно, указанная модель доступно по ссылке, см., например: https://github.com/bureaucratic-labs/dostoevsky. Так, при дополнительном анализе перефразированных стилизованных текстов, параметры mean_pos_score и meannegscore говорят о том, что с увеличе-

нием коэффициента стиля выраженность соответствующего стиля (Позитив/Негатив) повышается и становится достаточно ярко выраженной.

Модуль ранжирования 104 стилизованных перефразированных текстов может быть реализован на базе по меньшей мере одного вычислительного устройства, оснащенного соответствующим программным обеспечением для ранжирования и выбора лучшего кандидата. Так, например, в качестве критерия ранжирования может быть использовано расстояние Левенштейна. Так, в одном частном варианте осуществления, указанный модуль 104 выполнен с возможностью осуществления алгоритма вычисления попарных расстояний между исходным текстом и стилизованными текстами. Так, в еще одном частном варианте осуществления, сгенерированные модулем 103 кандидаты затем попадают в модуль ранжирования 104. В данном модуле для каждого кандидата вычисляется его близость к исходному предложению по метрике BertScore, которая оценивает смысловую близость исходного предложения и сгенерированного:

$$BertScore(original\ text, candidate_i), i = 1, \dots, n\_cadidates$$

Чем больше значение BertScore, тем ближе по смыслу парафразированный текст к исходному. В качестве финального варианта стилизованного перефразированного текста выбирается кандидат, для которого значение BertScore максимально:

$$k = \underset{i}{\operatorname{argmax}}\, BertScore(original\ text, candidate_i), i = 1, \dots, n\_cadidates$$

Кроме того, в еще одном частном варианте осуществления, ранжирование осуществляется на основе расстояния Левенштейна. Для этого вычисляются попарные расстояние между исходной нейтральной фразой и каждым из возможных кандидатов. В качестве лучшего выбирается кандидат, расстояние Левенштейна для которого минимально. Расстояние Левенштейна вычисляется с помощью функции distance из библиотеки Levenshtein. В результате получаем перефразированный стилизованный текст, с наименьшим изменением исходного текстового фрагмента, что соответственно повышает точность всего алгоритма перефразирования в заданном стиле и уменьшает вероятность добавления новых фактов.

Для специалиста в данной области техники будет очевидно, что могут применяться и другие методы ранжирования, известные из уровня техники и обеспечивающие возможность ранжирования текстовых фрагментов.

Модуль постобработки текста 105 может быть реализован на базе по меньшей мере одного вычислительного устройства. Модуль 105 выполнен с возможностью получения лучших отобранных модулем 104 вариантов для каждого текстового блока. Таким образом, если исходный текст был разбит на N текстовых фрагментов, то на вход модуль постобработки получает N стилизованных парафразов, таких что p_i, i=1,...,N - парафраз для исходного текстового фрагмента i. В данном модуле осуществляется постобработка парафразов: удаление дубляжа, спецсимволов и лишних символов, которые могли быть сгенерированы моделью. Данная обработка основана на правилах и регулярных выражениях. После этого происходит объединение стилизованных парафразов для отдельных текстовых фрагментов в единый парафраз для всего исходного текста. А именно полученные парафразы просто объединяются с сохранением порядка. Полученный объединённый текст и есть результат работы системы 100.

Для специалиста в данной области техники очевидно, что, хотя и описанные выше модули представлены как отдельные устройства, указанные модули также могут быть объединены в составе одного устройства, например, системы 300.

На фиг. 2 представлена блок схема способа 200 генерации текста в системе перефразирования текста, который раскрыт поэтапно более подробно ниже. Указанный способ 200 заключается в выполнении этапов, направленных на обработку различных цифровых данных. Обработка, как правило, выполняется с помощью системы, например, системы 100, которая также может представлять, например, сервер, компьютер, мобильное устройство, вычислительное устройство и т.д.

На этапе 210 система 100 получает входные данные, содержащие исходный текст на естественном языке. Так, входные данные могут быть получены от диалоговой системы по каналам передачи данных, таких как Интернет. Кроме того, исходный текстовый фрагмент может быть получен, непосредственно, с помощью интерфейса ввода/вывода системы 100. Так, текстовый фрагмент, в одном частном варианте осуществления, может содержать по меньшей мере два предложения. Так, в еще одном частном варианте осуществления, входной текстовый фрагмент может содержать десятки страниц. Форматом получаемых текстовых фрагментов может являться, например, текстовый формат файлов, такой как .txt. Текстовый фрагмент может быть загружен в интерфейс системы 100 и т.д.

На этапе 220 получают целевой стиль текстового фрагмента, характеризующий стилистические черты, присущие указанному целевому стилю, и параметр стилизации текста, характеризующий степень стилизации текста.

На этапе 220, система 100 также получает параметр требуемого стиля для перефразируемого текстового фрагмента и параметр стилизации текста, характеризующий степень стилизации текста. Так, в качестве параметра стиля, характеризующего стилистические черты, присущие указанному целевому стилю, может быть выбран один из заданных стилей, который может представлять определенные характерные черты речевого стиля, например, разговорный стиль общения, которому присущи, например, на-

личие неформальных обращений к пользователю, деловой стиль общения, которому присущи официальные обращения и т.д. Также, в одном частном варианте осуществления целевой стиль может указывать на эмоциональный окрас текста, например, грустный, веселый, нейтральный и т.д. Целевой стиль может быть определен настройками системы 100. Так, например, в одном частном варианте осуществления, целевой стиль может быть задан посредством GUI системы 100, например, с помощью указания в системе 100, требуемого стиля. В еще одном частном варианте осуществления, целевой стиль, например, научный, разговорный, деловой, может быть задан посредством всплывающих окон интерфейса системы 100. Для специалиста в данной области техники очевидно, что в качестве метода получения целевого стиля может быть использован любой известный из уровня техники метод взаимодействия пользователя с вычислительной техникой.

Кроме того, в дополнении к выбранному стилю, пользователь системы 100, такой как пользователь 110, также может задать параметр стилизации текста, характеризующий степень стилизации текста. Указанный параметр, в одном частном варианте осуществления, может быть задан одновременно с целевым стилем. В еще одном частном варианте осуществления, параметр стилизации текста может быть представлен в виде шкалы с уровнями градации (например от 1 до 10 и/или по степеням: слабая стилизация, умеренная стилизация, сильная стилизация). Параметр стилизации текста характеризует значение, определяющее требуемую степень стилизации текста. Так, на основе указанного параметра, модель МО изменяет (корректирует) исходные текстовые фрагменты (модуль 103) под требуемую степень стилизации текста.

На этапе 230 обрабатывают текстовый фрагмент, причем обработка включает по меньшей мере разбиение указанного фрагмента на текстовые блоки.

Так, на этапе 230, например, с помощью модуля 101, исходные текстовые данные, как в формате запроса пользователя, так и в формате текстового файла, обрабатывают для удаления из них спецсимволов, лишних пробелов, лишних повторов, и разбивают обработанные текста на отдельные текстовые блоки. Обработка текстовых данных может выполняться, например, посредством регулярных выражений. Кроме того, также могут применяться автоматические методы, основанные на регулярных выражениях и программно-аппаратные средства проверки пунктуации, например, Microsoft Office®, Google® и т.д.

Разбиение длинных текстов на отдельные текстовые блоков, может происходить на основе конца предложения и/или максимальной допустимой длины текстового блока, например, 200 слов. При разбиении, исходный текст разделяется на отдельные текстовые блоки. Разбиение происходит по предложениям так, чтобы в каждом блоке был законченный набор предложений (т.е. предложение посередине не разбивается, чтобы не получилось так, что половина попала в один фрагмент, а половина в другой). В случае, если произвести разбиение так, чтобы в каждом текстовом фрагменте оказался законченный набор предложений невозможно, то разбиение происходит по максимально допустимой длине блока.

В еще одном варианте осуществления, для ускорения работы системы допускается обработка текстовых фрагментов по батчам, то есть наборами (или пакетами) в определенном количестве.

На этапе 240 осуществляют кодирование текстовых блоков, полученных на этапе 230, причем в ходе кодирования выполняют по меньшей мере токенизацию текстовых блоков. Указанный этап 240 может выполняться модулем 102. Входной текст (текстовые блоки) может быть разделен на токены. Под токеном в данном решении следует понимать последовательность символов в тексте, которая имеет значение для анализа. В еще одном частном варианте осуществления токенизация текста может быть выполнена с помощью алгоритма BPE (Byte Pair encoding). В еще одном частном варианте осуществления токенизация может представлять собой разбиение текста на слова по пробелу между словами. Далее составляется словарь токенов фиксированного размера (например, 30000 токенов), где каждому токену сопоставляется его индекс в словаре. Пример токенизации на слова:

['*Вот что я нашел по вашей заявке*' → '*<Вот> <что> <я> <нашел> <по> <твоей><заявке>* ']

На указанном этапе 240 выполняется векторизация токенизированных текстов. Таким образом, токенизированный фрагмент текста (список токенов) после векторизации отображается в вектор индексов данных токенов в словаре. Приме векторизации при токенизации по словам:

'мама мыла раму' → *['мама', 'мыла', 'раму']* → *[235, 376, 1056]*

Далее способ 100 переходит к этапу 250.

На этапе 250 осуществляют обработку векторных представлений токенов текстовых блоков, с помощью модуля 103, в ходе которой осуществляется перефразирование текстового блока и его стилизация в заданном речевом стиле.

Так, на этапе 250 формируется набор кандидатов стилизованных перефразированных текстов для текстового блока в векторизированном виде. Для этого, осуществляется, например, посредством системы, такой как система перефразирования текста, генерация перефразированного текста из исходного текстового фрагмента, с помощью первой модели машинного обучения, обученной на перефразированных наборах данных, причем, в ходе генерации перефразированного текста, осуществляют перевзвешивание и корректировку первой модели, с учетом параметра стилизации текста, с помощью второй модели ма-

шинного обучения на базе нейронной сети, обученной на стилизованных наборах данных.

Как упоминалось выше, на этапе 250, принимают на вход закодированный текстовый фрагмент, стиль, в котором будет перефразирован текст и коэффициент выраженности стиля (коэффициент силы стиля) $w \geq 0$ (чем больше w, тем сильнее выражен стиль). Стоит отметить, что коэффициент силы стиля также может быть предварительно установлен в систему и может являться неизменным. Результатом выполнения шагов этапа 250 является набор кандидатов стилизованных перефразированных текстов для текстового блока в векторизированном виде.

Рассмотрим более подробно алгоритм взаимодействия подмодулей, входящих в состав модуля 103. На фиг. 3 представлена блок схема алгоритма перефразирования в заданном речевом стиле 300 текстового блока.

Одной из особенностей настоящего технического решения является возможность формирования перефразированного текста в заданном речевом стиле с высокой степенью оригинальности сгенерированного текста при сохранении исходного смыслового содержания. Указанные особенности, в частности, достигаются за счет реализации алгоритма 300.

Как упоминалось выше, обеспечение высокой степени оригинальности перефразированного текста и его стилизация в заданном речевом стиле является сложной и нетривиальной задачей. Для осуществления указанного процесса, в настоящем техническом решении на этапе 310 предсказывают вероятность следующего токена для каждого векторизованного текстового представления токенов текстового блока с помощью первой модели машинного обучения на базе нейронной сети.

Первая модель машинного обучения представляет собой модель машинного обучения на базе нейронной сети подмодуля перефразирования модуля 103, обученная на перефразированных текстах. Так, при появлении в модели исходного текста (закодированного текстового блока), за счет измененных весовых коэффициентов, наиболее вероятным предложением для модели будет рерайт (= парафраз) исходного текста (= исходный текст парафразированный другими словами), т.е. осуществляется повышение вероятности продолжения текстового фрагмента в том формате, в котором должен быть текст. Результатом выполнения этапа 310 являлась возможность предсказывания вероятности следующего токена на основе предыдущего начального фрагмента текста в перефразированном виде, т.е. генерация перефразированного текстового блока.

Стоит отметить, что такой подход обеспечивает возможность перефразирования всего текстового блока целиком, т.е. обеспечивает изменение конструкции текста с сохранением его смыслового содержания. Так, например, при изменении входного текстового блока, перефразированный блок может содержать как и синонимы для наиболее часто встречающихся слов, так и замены определенных конструкций в словосочетаниях, при сохранении смыслового содержания.

Пример: "Саша хороший сотрудник. Маша и Петя тоже хорошие сотрудники". Перефразированный текст: "Саша, Маша и Петя - отличные работники". Далее, на этапе 320 обрабатывают данные, полученные на этапе 310, с помощью второй модели машинного обучения на базе нейронной сети, в ходе которой осуществляют перевзвешивание и корректировку предсказания первой модели, в ходе которой перевзвешивают вероятности следующего токена, с учетом параметра стилизации текста. Так, на указанном этапе 320, выход первой модели машинного обучения поступает во вторую модель машинного обучения. Вторая модель машинного обучения реализована в подмодуле стилизации модуля 103 и дообучена на задачу классификации заданного набора стилей. Такой классификатор стилей затем используется для управления генерацией подмодуля перефразирования, чтобы направлять генерацию парафраза в русло заданного стиля. Так, в результате дообучения второй модели машинного обучения выполнялось настраивание весов обученной модели в соответствии с решаемой задачей. После подобного дообучения модель, за счет измененных весовых коэффициентов, для заданного текстового фрагмента способна предсказывать его стиль, например, позитивный, негативный, разговорный, официальный, грубый, книжный и другие.

Указанная вторая модель машинного обучения применялась для реализации стилизации в процессе парафразирования текстового блока, при этом при стилизации вторая модель может влиять на стилизацию текста в разных стилях. Т.е. одной из особенностей заявленного технического решения являлась возможность парафразирования текста с высокой точностью с одновременной стилизацией в одном из, заданного широкого набора, стилей, который осуществляется на основании корректировки вероятностей при генерации следующего токена и повышении вероятности генерации токена в заданном стиле. Указанный подход, в частности, позволяет снизить смысловое искажение текста за счет парафразирования и стилизации текста с использованием двух взаимодействующих моделей машинного обучения. Кроме того, при применении лишь одной модели машинного обучения для парафразирования и стилизации, высока вероятность генерации неоригинального текста. А при последовательном применении двух моделей машинного обучения (сначала парафразирование, а потом стилизация), а не в процессе, основанном на их взаимодействии, повышается вероятность искажения смысла исходного текста и нарушения фактов в указанном тексте.

Для устранения указанных проблем, а также обеспечения сохранения смыслового содержания перефразированного текста при его стилизации, на этапе 320 при генерации каждого следующего токена

выполнялась корректировка предсказаний модели парафразирования текста с учетом заданного целевого стиля и коэффициента выраженности стиля. Т.е. вероятности следующих токенов, предсказанные первой моделью, перевзвешивались с помощью второй модели, с учетом заданного стиля. Таким образом токены, соответствующие заданному стилю, получают большую вероятность. При этом, коэффициент выраженности стиля, управляет степенью того, какое влияние оказывает вторая модель и насколько сильно изменяет ее вероятности, придавая больший вес токенам заданного стиля.

Рассмотрим более подробно указанный этап 320.

Так, при генерации текста, подмодуль стилизации управляет генерацией подмодуля перефразирования текста следующим образом: на каждом шаге подмодуль перефразирования предсказывает вероятность следующего токена, а подмодуль стилизации осуществляет перевзвешивание предсказанных вероятностей следующего токена, с учетом заданного стиля, придавая большую вероятность токенам (и как следствие словам) заданного целевого стиля. Более детально, как было сказано в описании подмодуля перефразирования, дообученная модель способна предсказывать вероятность следующего токена на основе предыдущего начального фрагмента текста. Подмодуль стилизации способен оценивать вероятность класса (= стиля) для данного фрагмента текста. Таким образом, при стилевой генерации перефразированного текста вероятности следующего токена, полученные от подмодуля перефразирования перевзвешиваются с учетом вероятности целевого стиля от подмодуля стилизации следующим образом:

$$p_{final}(y_t|y_{1:t-1}, x, style) = p_{\text{перайтер}}(y_t|y_{1:t-1}, x)p_{style}(style|y_{1:t-1}, y_t)^w,$$

где - style - стиль текста, x - исходный парафразируемый текстовый фрагмент, y_t - предсказываемый следующий токен, $y_{1:t-1}$ - начальный префикс (отрезок) парафразированного текста, сгенерированный моделью, w - коэффициент силы стиля. Стоит отметить, что при w=0 стилевая модель не оказывает никакого эффекта ($p_{style}(style|y_{1:t-1},y_t)^0=1$), стилизации не происходит и текст перефразируется нейтрально. При этом, чем больше w тем более выраженным становится стиль, однако при очень больших значениях w, например, при w>20, подмодуль стилизации может оказывать слишком сильное влияние, что может повлечь искажение исходного смысла предложения. В связи с чем, в одном частном варианте осуществления, как уже упоминалось выше, коэффициент выраженности стиля (w) может выбираться пользователем из предопределенных значений и/или может быть интерпретирован в естественный язык, например, через указание степени стилизации.

Соответственно, после данной переоценки вероятностей, токены целевого стиля становятся более вероятными, а вероятность токенов, не соответствующих данному стилю понижается. Тем самым при генерации модуль 103 будет с большей вероятностью генерировать продолжение предложения в определенном стиле.

Продолжая алгоритм 300, на этапе 330 генерируют следующий токен для каждого векторизованного текстового представления токенов текстового блока на основе распределения вероятностей следующего токена, полученных на этапе 320, и добавляют указанный токен в конец векторизованного текстового представления токенов блока.

На указанном этапе 330, с учетом перевзвешенных и скорректированных предсказаний первой модели машинного обучения, выполняется последовательная генерация каждого следующего токена для текстового блока.

На этапе 340 генерируют векторизованное текстовое представление стилизованного перефразированного текста для текстового блока, итеративно повторяя этапы 310-330 до первого критерия останова.

Так, на указанном этапе 340 для всего текстового блока генерируется, посредством итеративного повторения этапов 310-330 до первого критерия останова, векторизованное представление стилизованного перефразированного текста для текстового блока. Так, первым критерием останова может являться максимальная допустимая длина текстового блока, например, 200 слов, 300 слов и т.д. Кроме того, в еще одном частном варианте осуществления, первый критерий останова представляет собой символ, соответствующий концу предложения. Т.е. на указанном этапе 340 осуществляют генерацию векторизованного текстового представления стилизованного перефразированного текста для текстового блока.

На этапе 350 генерируют по меньшей мере один стилизованный перефразированный текст для текстового блока в векторизированном виде итеративно повторяя этапы 310-340 до второго критерия останова.

Так, на указанном этапе 350, осуществляют генерацию множества стилизованных перефразированных кандидатов для исходного текстового блока, итеративно повторяя этапе 310-340 до второго критерия останова. Так, для каждого текстового блока генерируется набор кандидатов из которых, в дальнейшем, выбирается наиболее релевантный кандидат. Указанная особенность также позволяет обеспечить высокую семантическую точность перефразированных текстов за счет генерации нескольких кандидатов. Так, за счет генерации массива текстов обеспечивается возможность дальнейшей проверки и обработки всех вариантов перефразированного стилизованного текста для выбора наиболее семантически близкого к исходному тексту из массива. Кроме того, генерация именно массива перефразированных стилизованных текстов позволяет в дальнейшем исключить добавление новых фактов и/или некорректное изменение определённых слов на основе дальнейшей обработки таких текстов. При этом, второй критерий ос-

танова может представлять собой целочисленное значение, характеризующее требуемое количество кандидатов, например, четыре, пять и т.д. Указанный критерий может быть предопределен на стадии обучения модели и/или задан автоматически, например, в качестве входного критерия останова для способа 200. На этапе 360 формируют набор кандидатов стилизованных перефразированных текстов для текстового блока в векторизированном виде на основе данных, полученных на этапе 350.

Так, на этапе 360, весь набор сгенерированных кандидатов объединяется в единый файл для дальнейшего ранжирования.

Таким образом, за счет выполнения алгоритма 300, на этапе 250 осуществляют генерацию набора кандидатов стилизованных перефразированных текстов для текстового блока в векторизированном виде. Далее способ 200 переходит к этапу 260.

На этапе 260 сформированный набор кандидатов (массив) векторизованных перефразированных стилизованных текстов поступает в модуль 102. На указанном этапе 260 осуществляют декодирование каждого векторизованного перефразированного стилизованного текста из набора кандидатов, причем в ходе декодирования выполняют по меньшей мере преобразование кандидата в токены и детокенизацию. Так, например, в ходе указанного процесса каждый вектор фиксированной длины на основе его размерности сопоставляется токен по индексу словаря, что позволяет представить каждый вектор в виде токена. Процесс детокенизации является обратным процессом к токенизации и заключается в объединении токенов в текст. В результате выполнения данного этапа 260, набор кандидатов векторизованных перефразированных стилизованных текстов преобразуется в набор кандидатов перефразированных стилизованных текстов на естественном языке. Далее, на этапе 270 осуществляют ранжирование набора кандидатов стилизованного перефразированного текста и выбор лучшего перефразированного стилизованного кандидата, причем выбор лучшего перефразированного стилизованного кандидата основан на попарном расстоянии между исходным текстовым фрагментом блока и каждым из возможных перефразированных стилизованных текстовых фрагментов из набора.

Так, на этапе 270 перефразированные стилизованные тексты ранжируются по посимвольной близости с исходным текстом. Указанное ранжирование осуществляется на основе расстояния Левенштейна. Для этого вычисляются попарные расстояние между исходной нейтральной фразой и каждым из возможных кандидатов. В качестве лучшего выбирается кандидат, расстояние Левенштейна для которого минимально. Расстояние Левенштейна вычисляется с помощью функции distance из библиотеки Levenshtein. В результате получаем перефразированный стилизованный текст, с наименьшим изменением исходного текстового фрагмента, что соответственно повышает точность всего алгоритма перефразирования в заданном стиле и уменьшает вероятность добавления новых фактов. Кроме того, как указывалось выше, ранжирование также может быть выполнено с помощью метрики BertScore.

На этапе 280 выполняют объединение стилизованных перефразированных текстов каждого блока с сохранением исходного порядка в перефразированный стилизованный текстовый фрагмент и отправляют указанный фрагмент в систему перефразирования текста.

Поскольку, как указывалось выше, текстовый фрагмент может быть достаточно большой длины, то на указанном этапе 280, осуществляется объединение, с сохранением исходного порядка, перефразированных стилизованных текстовых фрагментов в единый текстовый фрагмент. Так, в одном частном варианте осуществления, на указанном этапе 280 каждый выбранный кандидат объединяется с другими выбранными кандидатами. Для наглядности, рассмотрим пример, структурно показывающий принцип обработки и объединения текстовых фрагментов.

Так, если входной текстовый фрагмент содержит в себе 10 предложений, то на этапе 230 указанный фрагмент будет разбит на несколько текстовых блоков (например, два блока), соответствующих максимальной длине блока (например по 200 слов или 5 предложений), причем блоки маркируются последовательно. Далее, каждый текстовый блок, например, параллельно, обрабатывается в соответствии с этапами способа 240-270. На этапе 280 в соответствии с маркировкой блока (позиции блока), объединяются перефразированные стилизованные текста из блока. Т.е. формируется итоговый перефразированный стилизованный текст для всего исходного текста. Так, в одном частном варианте осуществления, итоговый сгенерированный перефразированный стилизованный текстовый фрагмент отправляют в систему перефразирования текста.

Как указывалось выше, система перефразирования текста, например, система 100 может использоваться в таких сферах обработки текста, как: редактирование, подготовка новостных статей, презентаций, подготовка текста для цифровых ассистентов, диалоговых систем и т.д. Система 100 может быть интегрирована в процессы создания текстового контента организациями и обеспечивать изложение текстовых данных в едином уникальном стиле.

Соответственно, в одном частном варианте осуществления, текст, полученный на этапе 280, может быть сохранен в памяти системы, например, системы 100, в виде файла данных и/или отправлен, непосредственно, в интерфейс входа/выхода указанной системы 100. Кроме того, в еще одном частном варианте осуществления, итоговый текст может быть отправлен, по каналам передачи данных, для последующего отображения пользователю, например, с помощью интерфейса ввода-вывода пользовательского вычислительного устройства.

Так, в еще одном частном варианте осуществления, с помощью системы перефразирования текста может быть реализован способ перефразирования текста.

Для этого, на первом этапе, система перефразирования текста, получает запрос пользователя на перефразирование текстового фрагмента на естественном языке и целевой стиль текста. Кроме того, в еще одном частном варианте осуществления, к указанным входным данным дополнительно может быть отправлен и коэффициент выраженности стиля.

На следующем этапе, система перефразирования текста, осуществляет генерацию перефразированного стилизованного текста из исходного в соответствии с этапами способа 200.

После этого, сгенерированный текст отображается, например, в качестве ответа на запрос пользователя. Как будет очевидно, метод получения пользователем итогового сгенерированного перефразированного текста может также представлять, например, получение файла данных, содержащий итоговый сгенерированный текст, получение указанного текста в диалоговой системе, т.е. в интерфейсе ввода/вывода как пользователя, так и непосредственно системы 100.

Таким образом, в вышеприведенных материалах были описаны система и способ генерации текста в системах перефразирования текста, обеспечивающий высокую семантическую точность генерации перефразированного стилизованного текста с высокой степенью оригинальности и возможностью получения текста с разной выраженностью стиля.

Кроме того, стоит отметить, что благодаря реализации заявленного решения также обеспечивается универсальность стилизации перефразированного текста, позволяющая генерировать текст не только в одном стиле, но и обеспечивать возможность стилизации исходного текста в нескольких стилях, в зависимости от сферы применения. Указанная особенность исключает необходимость в отдельном формировании для каждого стиля уникальной стилизованной реплики.

На фиг. 4 представлен пример общего вида вычислительной системы (400), которая обеспечивает реализацию заявленного способа или является частью компьютерной системы, например, модулями 101-105, сервером, персональным компьютером, частью вычислительного кластера, обрабатывающим необходимые данные для осуществления заявленного технического решения.

В общем случае система (400) содержит такие компоненты, как: один или более процессоров (401), по меньшей мере одну память (402), средство хранения данных (403), интерфейсы ввода/вывода (404), средство В/В (405), средство сетевого взаимодействия (306), которые объединяются посредством универсальной шины.

Процессор (401) выполняет основные вычислительные операции, необходимые для обработки данных при выполнении способа (200). Процессор (401) исполняет необходимые машиночитаемые команды, содержащиеся в оперативной памяти (402). Память (402), как правило, выполнена в виде ОЗУ и содержит необходимую программную логику, обеспечивающую требуемый функционал.

Средство хранения данных (403) может выполняться в виде HDD, SSD дисков, рейд массива, флэш-памяти, оптических накопителей информации (CD, DVD, MD, Blue-Ray дисков) и т.п. Средства (403) позволяют выполнять долгосрочное хранение различного вида информации, например сгенерированных стилизованных перефразированных текстов, идентификаторов пользователей, идентификаторов цифровых ассистентов и т.п. Для организации работы компонентов системы (400) и организации работы внешних подключаемых устройств применяются различные виды интерфейсов В/В (404). Выбор соответствующих интерфейсов зависит от конкретного исполнения вычислительного устройства, которые могут представлять собой, не ограничиваясь: PCI, AGP, PS/2, IrDa, FireWire, LPT, COM, SATA, IDE, Lightning, USB (2.0, 3.0, 3.1, micro, mini, type C), TRS/Audio jack (2.5, 3.5, 6.35), HDMI, DVI, VGA, Display Port, RJ45, RS232 и т.п. Выбор интерфейсов (404) зависит от конкретного исполнения системы (300), которая может быть реализована на базе широко класса устройств, например, персональный компьютер, мейнфрейм, ноутбук, серверный кластер, тонкий клиент, смартфон, сервер и т.п.

В качестве средств В/В данных (405) может использоваться: клавиатура, джойстик, дисплей (сенсорный дисплей), монитор, сенсорный дисплей, тачпад, манипулятор мышь, световое перо, стилус, сенсорная панель, трекбол, динамики, микрофон, средства дополненной реальности, оптические сенсоры, планшет, световые индикаторы, проектор, камера, средства биометрической идентификации (сканер сетчатки глаза, сканер отпечатков пальцев, модуль распознавания голоса) и т.п. Средства сетевого взаимодействия (406) выбираются из устройств, обеспечивающий сетевой прием и передачу данных, например, Ethernet карту, WLAN/Wi-Fi модуль, Bluetooth модуль, BLE модуль, NFC модуль, IrDa, RFID модуль, GSM модем и т.п. С помощью средств (405) обеспечивается организация обмена данными между, например, системой (400), представленной в виде сервера и вычислительным устройством пользователя, на котором могут отображаться полученные данные (сгенерированная стилизованная реплика цифрового ассистента) по проводному или беспроводному каналу передачи данных, например, WAN, PAN, ЛВС (LAN), Интранет, Интернет, WLAN, WMAN или GSM.

Конкретный выбор элементов системы (400) для реализации различных программно-аппаратных архитектурных решений может варьироваться с сохранением обеспечиваемого требуемого функционала.

Представленные материалы заявки раскрывают предпочтительные примеры реализации технического решения и не должны трактоваться как ограничивающие иные, частные примеры его воплощения,

не выходящие за пределы испрашиваемой правовой охраны, которые являются очевидными для специалистов соответствующей области техники. Таким образом, объем настоящего технического решения ограничен только объемом прилагаемой формулы.

ФОРМУЛА ИЗОБРЕТЕНИЯ

1. Способ генерации текста, выполняемый по меньшей мере одним вычислительным устройством и содержащий этапы, на которых:

a) получают текстовый фрагмент на естественном языке;

b) получают целевой стиль текстового фрагмента, характеризующий стилистические черты, присущие указанному целевому стилю, и параметр стилизации текста, характеризующий степень стилизации текста;

c) обрабатывают текстовый фрагмент, причем обработка включает по меньшей мере разбиение указанного фрагмента на текстовые блоки;

d) осуществляют кодирование каждого текстового блока, причем в ходе кодирования выполняют токенизацию текстового блока;

e) выполняют векторизацию текстовых блоков по токенам, полученных на этапе d);

f) осуществляют обработку данных, полученных на этапе e), в ходе которой:

i) осуществляют генерацию перефразированного текста из исходного текстового фрагмента, с помощью первой модели машинного обучения, обученной на перефразированных наборах данных, причем, в ходе генерации перефразированного текста, осуществляют перевзвешивание и корректировку первой модели, с учетом параметра стилизации текста, с помощью второй модели машинного обучения на базе нейронной сети, обученной на стилизованных наборах данных;

ii) формируют набор кандидатов стилизованных перефразированных текстов для текстового блока в векторизированном виде на основе данных, полученных на шаге i;

g) осуществляют декодирование каждого кандидата, полученного на этапе f), в каждом текстовом блоке, причем в ходе декодирования выполняют по меньшей мере преобразование векторизированных стилизованных текстов в токены и детокенизацию;

h) осуществляют ранжирование набора кандидатов стилизованного перефразированного текста и выбор лучшего перефразированного стилизованного кандидата, причем выбор лучшего перефразированного стилизованного кандидата основан на попарном расстоянии между исходным текстовым фрагментом блока и каждым из возможных перефразированных стилизованных текстовых фрагментов из набора;

i) объединяют стилизованные перефразированные тексты каждого блока с сохранением исходного порядка в перефразированный стилизованный текстовый фрагмент и отправляют указанный фрагмент в систему перефразирования текста.

2. Способ по п.1, характеризующийся тем, что текстовый блок не превышает заранее заданный параметр длины блока.

3. Способ по п.2, характеризующийся тем, что текстовый фрагмент разбивается на текстовые блоки не превышающие заданный параметр длины блока.

4. Способ по п.1, характеризующийся тем, что текстовый фрагмент содержит по меньшей мере два предложения.

5. Способ по п.1, характеризующийся тем, что текстовые блоки состоят из законченного числа предложений.

6. Способ по п.1, характеризующийся тем, что, в ходе генерации перефразированного текста из исходного текстового фрагмента, выполняют следующие шаги:

i) предсказывают вероятность следующего токена для каждого векторизованного текстового представления токенов текстового блока с помощью первой модели машинного обучения на базе нейронной сети;

ii) обрабатывают данные, полученные на шаге i, с помощью второй модели машинного обучения на базе нейронной сети, в ходе которой осуществляют перевзвешивание и корректировку предсказания первой модели, в ходе которой перевзвешивают вероятности следующего токена, с учетом параметра стилизации текста;

iii) генерируют следующий токен для каждого векторизованного текстового представления токенов текстового блока на основе распределения вероятностей следующего токена, полученных на шаге ii, и добавляют указанный токен в конец векторизованного текстового представления токенов блока;

iv) генерируют векторизованное текстовое представление стилизованного перефразированного текста для текстового блока, итеративно повторяя шаги i-iii до первого критерия останова;

v) генерируют по меньшей мере один стилизованный перефразированный текст, для текстового блока в векторизированном виде итеративно повторяя шаги i-iv до второго критерия останова.

7. Способ по п.6, характеризующийся тем, что первый критерий останова представляет собой максимальную допустимую длину текстового блока или символ, соответствующий концу предложения.

8. Способ по п.7, характеризующийся тем, что допустимая длина текстового блока не превышает

200 слов.

9. Способ по п.6, характеризующийся тем, что второй критерий останова представляет собой целочисленное значение, характеризующее требуемое количество кандидатов.

10. Система генерации текста, содержащая:

по меньшей мере один процессор;

по меньшей мере одну память, соединенную с процессором, которая содержит машиночитаемые инструкции, которые при их выполнении по меньшей мере одним процессором обеспечивают выполнение способа по любому из пп.1-9.

11. Способ перефразирования текста, выполняемый по меньшей мере одним вычислительным устройством и содержащий этапы, на которых:

a) получают запрос пользователя на перефразирование текстового фрагмента на естественном языке и целевой стиль текста;

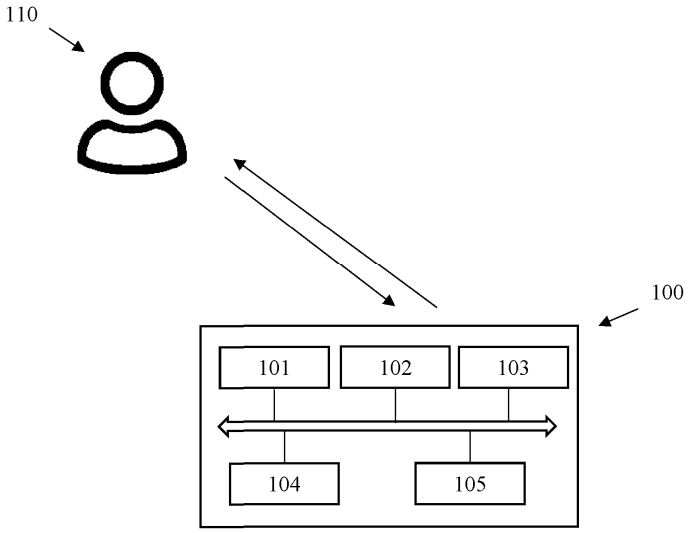
b) обрабатывают, полученный на этапе a), текстовый фрагмент, с помощью способа по любому из пп.1-9;

c) отображают ответ на запрос пользователя, содержащий перефразированный стилизованный текстовый фрагмент, полученный на этапе b).

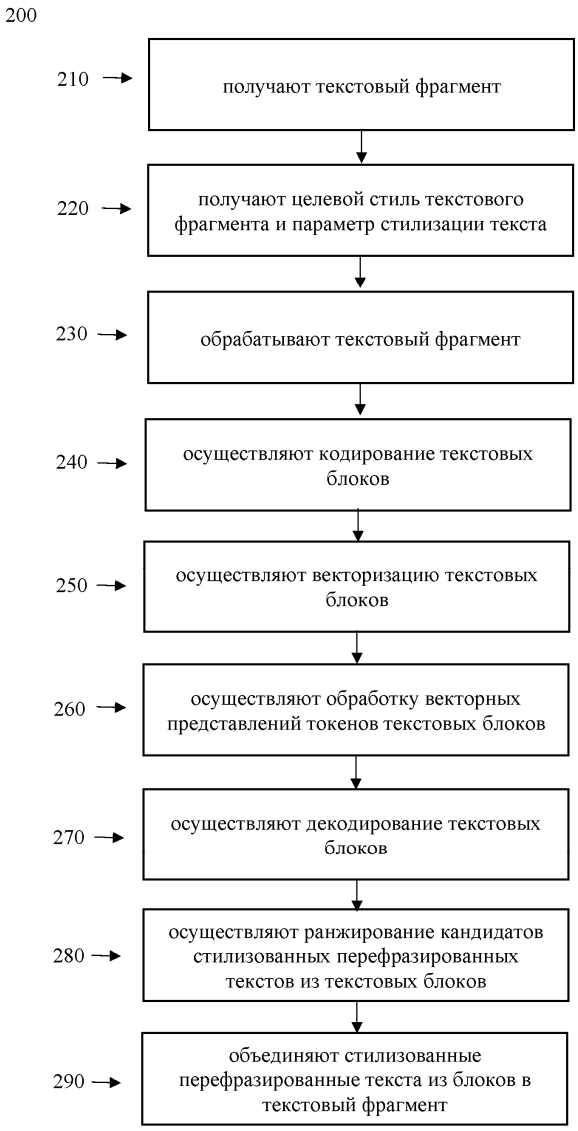
12. Система перефразирования текста, содержащая:

по меньшей мере один процессор;

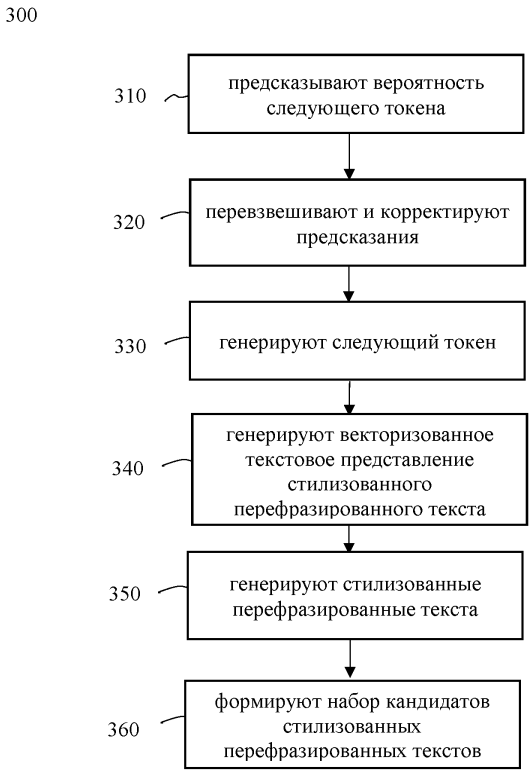
по меньшей мере одну память, соединенную с процессором, которая содержит машиночитаемые инструкции, которые при их выполнении по меньшей мере одним процессором обеспечивают выполнение способа по п.11.



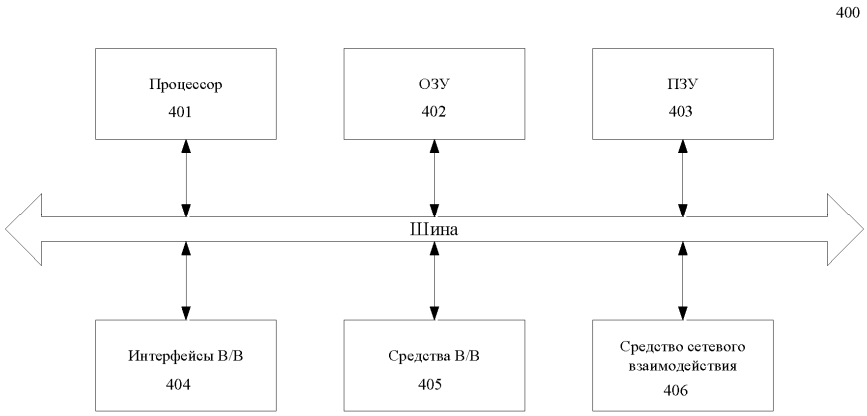
Фиг. 1


200
210
получают текстовый фрагмент
220
получают целевой стиль текстового фрагмента и параметр стилизации текста
230
обрабатывают текстовый фрагмент
240
осуществляют кодирование текстовых блоков
250
осуществляют векторизацию текстовых блоков
260
осуществляют обработку векторных представлений токенов текстовых блоков
270
осуществляют декодирование текстовых блоков
280
осуществляют ранжирование кандидатов стилизованных перефразированных текстов из текстовых блоков
290
объединяют стилизованные перефразированные текста из блоков в текстовый фрагмент

Фиг. 2


300
310
предсказывают вероятность следующего токена
320
перевзвешивают и корректируют предсказания
330
генерируют следующий токен
340
генерируют векторизованное текстовое представление стилизованного перефразированного текста
350
генерируют стилизованные перефразированные текста
360
формируют набор кандидатов стилизованных перефразированных текстов

Фиг. 3


400
Процессор
401
ОЗУ
402
ПЗУ
403
Шина
Интерфейсы В/В
404
Средства В/В
405
Средство сетевого взаимодействия
406

Фиг. 4