

ФЕДЕРАЛЬНАЯ СЛУЖБА
ПО ИНТЕЛЛЕКТУАЛЬНОЙ СОБСТВЕННОСТИ

(12) ОПИСАНИЕ ИЗОБРЕТЕНИЯ К ПАТЕНТУ

(52) СПК

G06N 3/08 (2023.05)

(21)(22) Заявка: 2020132862, 06.10.2020

(24) Дата начала отсчета срока действия патента:
06.10.2020Дата регистрации:
21.06.2023

Приоритет(ы):

(22) Дата подачи заявки: 06.10.2020

(43) Дата публикации заявки: 25.04.2022 Бюл. № 12

(45) Опубликовано: 21.06.2023 Бюл. № 18

Адрес для переписки:

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

(72) Автор(ы):

Петров Алексей Сергеевич (RU),
Губанов Сергей Дмитриевич (RU),
Гайдаенко Сергей Александрович (RU)

(73) Патентообладатель(и):

ОБЩЕСТВО С ОГРАНИЧЕННОЙ
ОТВЕТСТВЕННОСТЬЮ «ЯНДЕКС» (RU)(56) Список документов, цитированных в отчете
о поиске: RU 2731335 C2, 01.09.2020. JP
2020038687 A, 12.03.2020. US 10572598 B2,
25.02.2020. CN 109977220 A, 05.07.2019. CN
109885683 A, 14.06.2019. CN 110750240 A,
04.02.2020. US 2009/0187516 A1, 23.07.2009. CN
110134782 A, 16.08.2019.(54) СПОСОБ И СЕРВЕР ДЛЯ ОБУЧЕНИЯ НЕЙРОННОЙ СЕТИ ФОРМИРОВАНИЮ ТЕКСТОВОЙ
ВЫХОДНОЙ ПОСЛЕДОВАТЕЛЬНОСТИ

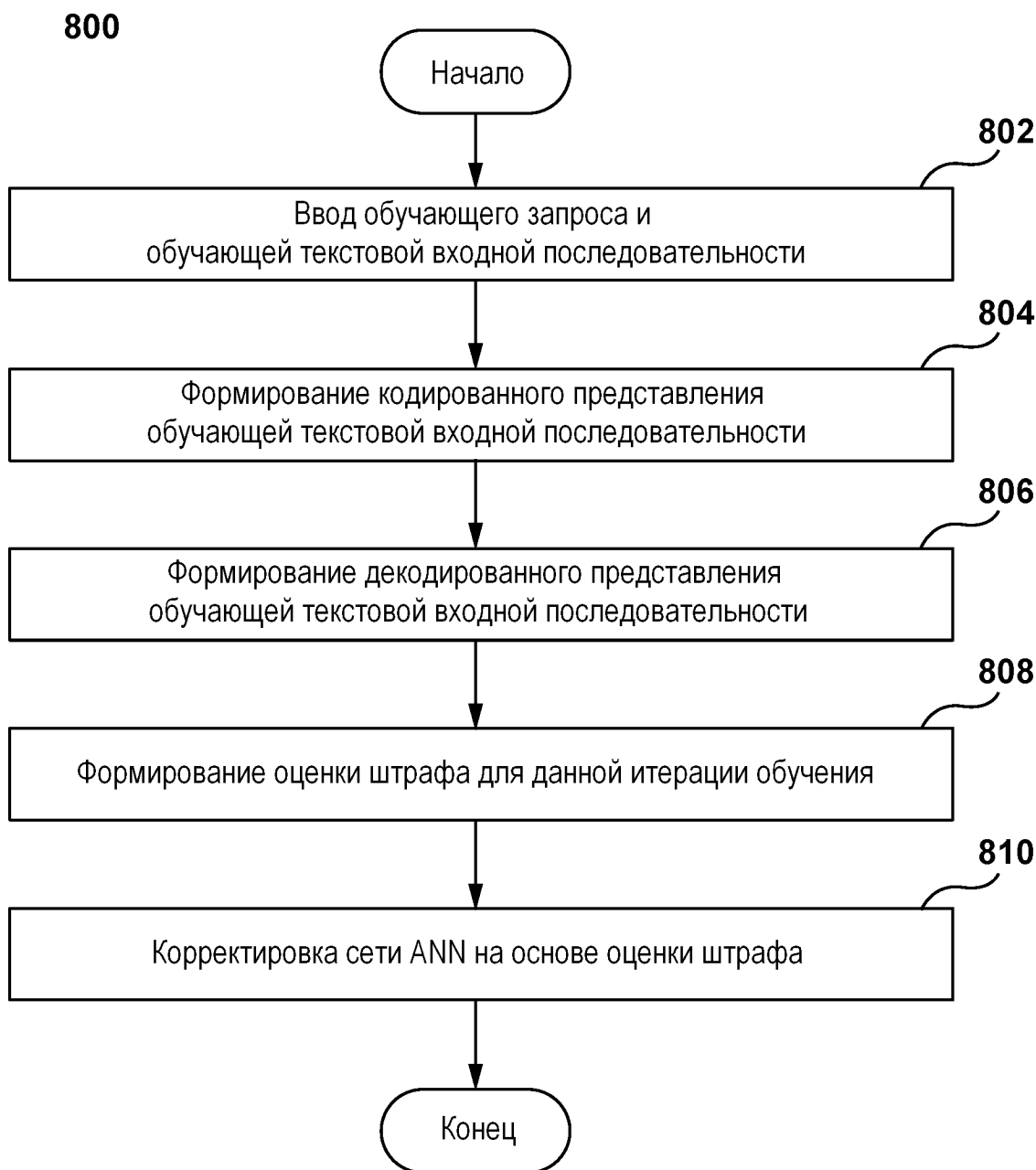
(57) Реферат:

Изобретение относится к способу и серверу для обучения нейронной сети внимания (ANN) формированию текстовой выходной последовательности. Технический результат заключается в возможности формирования релевантной текстовой выходной последовательности, представляющей собой сводку контента множества контент-ресурсов. В способе осуществляют ввод сервером обучающего запроса и обучающей текстовой входной последовательности в подсеть кодера, при этом обучающая текстовая входная последовательность (а) формируется в виде последовательности обучающих сниппетов контента и (б) разделяется на последовательность входных групп, а входная группа связана с обучающим сниппетом контента и содержит слова из этого обучающего сниппета контента; формирование сервером, использующим подсеть кодера, кодированного представления обучающей

текстовой входной последовательности, включая формирование выходных данных вида «внимание» для соответствующих слов из обучающей текстовой входной последовательности путем применения маски ограничения внимания в отношении обучающей текстовой входной последовательности, а выходные данные вида «внимание» используются для формирования кодированного представления обучающей текстовой входной последовательности, при формировании выходных данных вида «внимание» для слова из входной группы маска ограничения внимания позволяет «принимать во внимание» только слова из этой входной группы так, чтобы выходные данные вида «внимание» формировались на основе контекста этой входной группы, а не контекстов других входных групп в обучающей текстовой входной последовательности; формирование сервером, использующим подсеть

декодера, декодированного представления для обучающей текстовой входной последовательности, соответствующего прогнозируемой текстовой выходной последовательности; формирование сервером оценки штрафа для итерации обучения путем

сравнения прогнозируемой текстовой выходной последовательности с заранее заданной текстовой выходной последовательностью, представляющей собой заранее заданный ответ на обучающий запрос; и корректировку сервером сети ANN на основе оценки штрафа. 2 н. и 30 з.п. ф-лы, 7 ил.



Фиг. 7



FEDERAL SERVICE
FOR INTELLECTUAL PROPERTY

(12) **ABSTRACT OF INVENTION**

(52) CPC

G06N 3/08 (2023.05)(21)(22) Application: **2020132862, 06.10.2020**(24) Effective date for property rights:
06.10.2020Registration date:
21.06.2023

Priority:

(22) Date of filing: **06.10.2020**(43) Application published: **25.04.2022** Bull. № 12(45) Date of publication: **21.06.2023** Bull. № 18

Mail address:

**119021, Moskva, ul. Lva Tolstogo, 16, OOO
"Yandex", patentnyj otdel, Malakhova Nadiya
Akhavovna**

(72) Inventor(s):

**Petrov Aleksey Sergeevich (RU),
Gubanov Sergey Dmitrievich (RU),
Gaydaenko Sergey Aleksandrovich (RU)**

(73) Proprietor(s):

**OBSheESTVO S OGRANICHENNOI
OTVETSTVENNOSTIu «IaNDEKS» (RU)**

(54) **METHOD AND SERVER FOR TEACHING A NEURAL NETWORK TO FORM A TEXT OUTPUT SEQUENCE**

(57) Abstract:

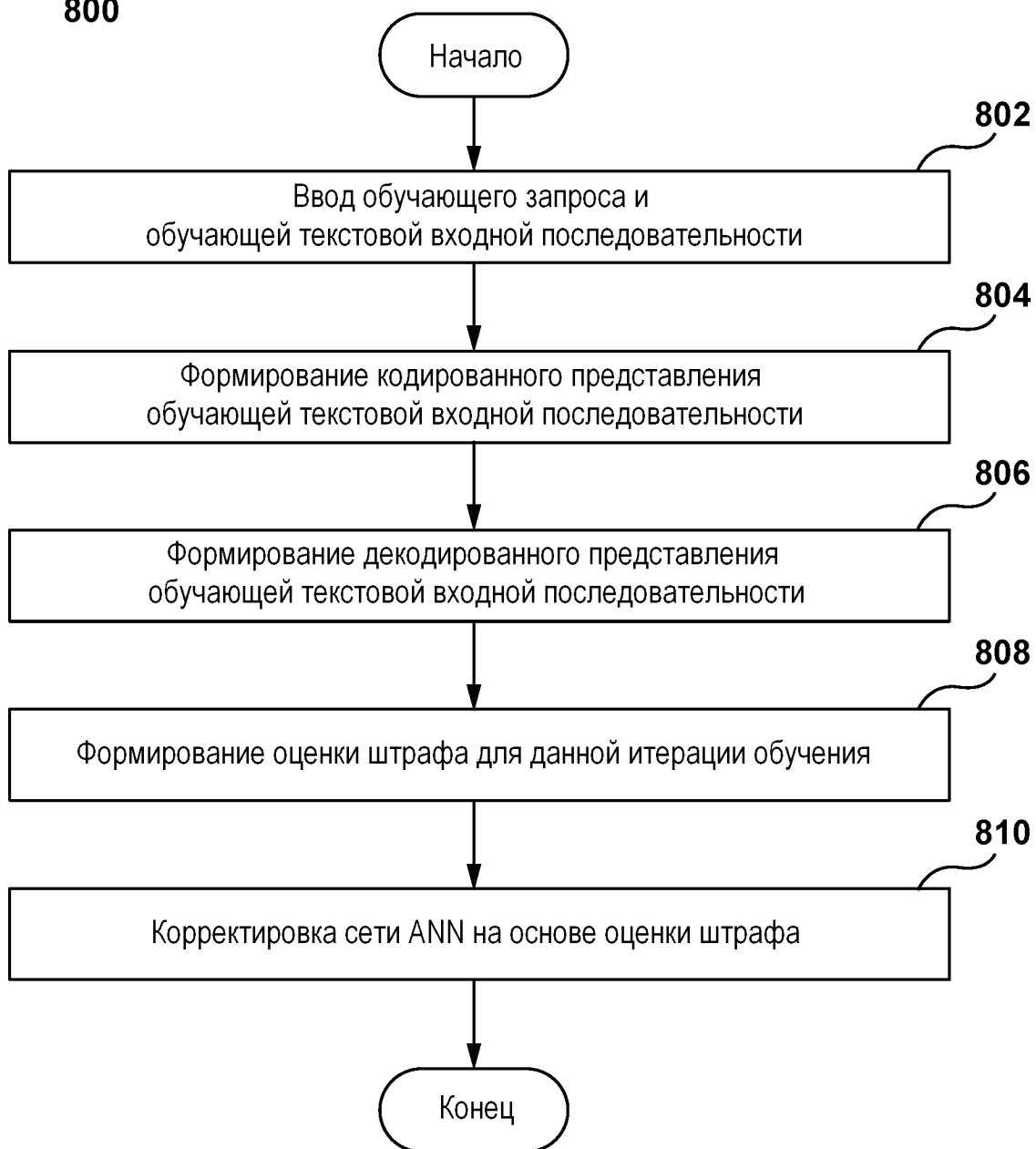
FIELD: neural networks.

SUBSTANCE: invention relates to a method and a server for training an attention neural network (ANN) to generate a text output sequence. The technical result consists in the possibility of generating a relevant text output sequence, which is a summary of the content of a plurality of content resources. In the method, the server inputs a training request and a training text input sequence into the encoder subnet, while the training text input sequence (a) is formed as a sequence of training content snippets and (b) is divided into a sequence of input groups, and the input group is associated with a training content snippet and contains words from this learning content snippet; the encoded representation of the training text input sequence is generated by the server using the encoder subnet, including the generation of the attention output for the corresponding words from the training text input sequence by applying an attention restriction mask to the training text input sequence, and the attention output

is used to form an encoded representation of the training text input sequence, when generating output data of the "attention" type for a word from the input group, the attention restriction mask allows you to "take into account" only words from this input group so that the output data of the "attention" type is formed based on the context this input group, and not the contexts of other input groups in the training text input sequence; generating, by the server using the decoder subnet, a decoded representation for the training text input sequence corresponding to the predicted text output sequence; generating, by the server, a penalty estimate for the training iteration by comparing the predicted text output sequence with a predetermined text output sequence representing a predetermined response to the training request; and adjusting the ANN network server based on the penalty estimate.

EFFECT: possibility of generating a relevant text output sequence, which is a summary of the content of a plurality of content resources.

800



Фиг. 7

Область техники, к которой относится изобретение

[01] Настоящая технология относится к системам интеллектуальных персональных помощников (IPA, Intelligent Personal Assistant) и, в частности, к способу и электронному устройству для обучения нейронной сети формированию текстовой выходной последовательности.

Уровень техники

[02] Электронные устройства, такие как смартфоны и планшеты, обеспечивают доступ к постоянно растущему количеству разнообразных приложений и сервисов для обработки информации различных видов и/или для доступа к ней. Тем не менее начинающие пользователи и/или пользователи с ограниченными возможностями и/или пользователи, управляющие транспортным средством, могут не иметь возможности эффективно взаимодействовать с такими устройствами, главным образом, вследствие разнообразия функций, предоставляемых такими устройствами, или из-за невозможности использовать интерфейсы пользователь-машина, предоставляемые такими устройствами (такие как клавиатура или сенсорный экран). Например, пользователь, управляющий транспортным средством, или пользователь с нарушениями зрения может не иметь возможности использовать клавиатуру на сенсорном экране, реализованную в некоторых из этих устройств.

[03] Для выполнения функций по запросам пользователя разработаны системы IPA. Например, системы IPA могут использоваться для поиска информации и для навигации. Традиционная система IPA, например, такая как система IPA Siri[®], может принимать от устройства простые вводимые человеком данные на естественном языке и выполнять для пользователя множество задач. Например, пользователь может обмениваться информацией с системой IPA Siri[®], предоставляя речевые фрагменты (через голосовой интерфейс Siri[®]), в частности, чтобы узнать текущую погоду, местоположение ближайшего торгового центра и т.п. Пользователь также может запрашивать выполнение различных приложений, установленных на электронном устройстве.

[04] Традиционные системы IPA способны обеспечивать пользователям возможность ввода речевых поисковых запросов. Например, традиционные системы IPA могут формировать текстовое представление произнесенного запроса и запускать поиск релевантной информации.

Раскрытие изобретения

[05] Системы IPA используются для поддержки речевого взаимодействия между пользователями и электронными устройствами. Это обеспечивает удобное взаимодействие пользователей с такими устройствами, например, когда пользователь имеет нарушения зрения, а также когда в устройстве не предусмотрен традиционный визуальный дисплей (например, в умной колонке) и/или когда руки или глаза пользователя заняты (например, при вождении).

[06] Разработчики настоящей технологии обнаружили определенные технические недостатки, связанные с существующими системами IPA. По меньшей мере один недостаток традиционных систем IPA заключается в их ограниченных возможностях работы с речевыми запросами и, в частности, в ограничениях при предоставлении ответов в звуковом формате. Это главным образом объясняется тем, что пользователи, ожидающие ответ в звуковом формате, обычно предпочитают краткие ответы, а разработка системы IPA, способной сжато предоставлять информацию в ответ множество запросов, которые пользователи потенциально могут вводить, представляет собой технически сложную задачу. Например, некоторые традиционные системы

непригодны для формирования сводки релевантного контента, достаточно краткой для предоставления пользователю в звуковом формате. В другом примере другие традиционные системы непригодны для использования контекста, к которому относится контент, для формирования лучшей сводки контента.

5 [07] Целью настоящего изобретения является устранение по меньшей мере некоторых недостатков известных решений. Разработчики настоящей технологии разработали систему, в которой некоторый контент, определенный как релевантный запросу, обрабатывается нейронной сетью (NN, Neural Network). В по меньшей мере некоторых не имеющих ограничительного характера вариантах осуществления настоящей
10 технологии сеть NN может представлять собой сеть NN вида «внимание» (ANN, Attention-type NN), содержащую подсеть кодера и подсеть декодера, способные обрабатывать последовательности данных, при этом в подсети кодера реализован конкретный механизм внимания.

[08] В по меньшей мере некоторых не имеющих ограничительного характера
15 вариантах осуществления настоящей технологии можно сказать, что механизм внимания подсети кодера способен применять маску ограничения внимания в отношении входной последовательности данных. Предполагается, что входная последовательность данных может содержать входные группы, соответствующие снippetам контента из релевантного запросу контента, полученного, например, от сервера поисковой системы,
20 способного выполнять веб-поиски в ответ на пользовательские поисковые запросы. В результате применения маски ограничения внимания в отношении элемента входной последовательности, в которой размечены такие входные группы, подсеть кодера может «принимать во внимание» контекст входной группы этого элемента, а не контекст других входных групп. Таким образом, можно сказать, что контекст, полученный для
25 элемента с помощью механизма внимания, основан на контексте входной группы этого элемента, а не на других контекстах других входных групп.

[09] Согласно первому аспекту настоящей технологии реализован способ обучения сети ANN формированию текстовой выходной последовательности на основе текстовой входной последовательности и запроса. Текстовая входная последовательность
30 формируется в виде последовательности снippetов контента, полученных из соответствующих контент-ресурсов, релевантных запросу. Текстовая выходная последовательность представляет собой сводку контента последовательности снippetов контента, которая должна использоваться в качестве ответа на запрос. Сеть ANN содержит подсеть кодера и подсеть декодера. Способ выполняется сервером, на котором
35 реализована сеть ANN. Способ на итерации обучения сети ANN включает в себя ввод сервером обучающего запроса и обучающей текстовой входной последовательности в подсеть кодера. Обучающая текстовая входная последовательность (а) формируется в виде последовательности обучающих снippetов контента и (б) разделяется на последовательность входных групп. Входная группа связана с обучающим снippetом
40 контента и содержит слова из обучающего снippetа контента. Способ на итерации обучения сети ANN включает в себя формирование сервером, использующим подсеть кодера, кодированного представления обучающей текстовой входной последовательности, включая формирование выходных данных вида «внимание» для соответствующих слов из обучающей текстовой входной последовательности путем
45 применения маски ограничения внимания в отношении обучающей текстовой входной последовательности. Выходные данные вида «внимание» используются для формирования кодированного представления обучающей текстовой входной последовательности. При формировании выходных данных вида «внимание» для слова

из входной группы маска ограничения внимания позволяет «принимать во внимание» только слова из этой входной группы так, чтобы выходные данные вида «внимание» формировались на основе контекста этой входной группы, а не контекстов других входных групп в обучающей текстовой входной последовательности. Способ на итерации обучения сети ANN включает в себя формирование сервером, использующим подсеть декодера, декодированного представления обучающей текстовой входной последовательности. Такое декодированное представление соответствует прогнозируемой текстовой выходной последовательности. Способ на итерации обучения сети ANN включает в себя формирование сервером оценки штрафа для итерации обучения путем сравнения прогнозируемой текстовой выходной последовательности с заранее заданной текстовой выходной последовательностью. Заранее заданная текстовая выходная последовательность представляет собой заранее заданный ответ на обучающий запрос. Способ на итерации обучения сети ANN включает в себя корректировку сервером сети ANN на основе оценки штрафа.

[10] В некоторых вариантах осуществления способа формируемые выходные данные вида «внимание» дополнительно основываются на контексте из обучающего запроса.

[11] В некоторых вариантах осуществления способа на итерации обучения сеть ANN обучается на основе обучающего набора данных, содержащего (а) обучающий запрос, (б) обучающую текстовую входную последовательность и (в) заранее заданную текстовую выходную последовательность. Способ на другой итерации обучения дополнительно включает в себя обучение сети ANN на основе другого обучающего набора данных. Другой обучающий набор данных содержит (а) обучающий запрос, (б) модифицированную обучающую текстовую входную последовательность и (в) заранее заданную текстовую выходную последовательность. Входные группы в модифицированной обучающей текстовой входной последовательности расположены в порядке, отличном от порядка входных групп в обучающей текстовой входной последовательности.

[12] В некоторых вариантах осуществления способ дополнительно включает в себя формирование обучающего снippets контента из последовательности обучающих снippets контента. Такое формирование включает в себя ввод сервером обучающего запроса и контента из обучающего контент-ресурса в алгоритм машинного обучения (MLA, Machine Learning Algorithm), способный выдавать прогнозируемый снippet контента. Алгоритм MLA обучен формированию прогнозируемого снippets контента так, чтобы прогнозируемый снippet контента был подобен заранее заданной текстовой выходной последовательности. Прогнозируемый снippet контента представляет собой обучающий снippet контента.

[13] В некоторых вариантах осуществления способа на итерации обучения сеть ANN обучается на основе обучающего набора данных, содержащего (а) обучающий запрос, (б) обучающую текстовую входную последовательность и (в) заранее заданную текстовую выходную последовательность. Итерация обучения выполняется на первом этапе обучения сети ANN, включающем в себя первое множество итераций обучения. Способ дополнительно включает в себя выполнение второго этапа обучения сети ANN. Второй этап обучения сети ANN содержит второе множество итераций обучения. На другой итерации обучения из второго множества итераций обучения сеть ANN обучается на основе другого обучающего набора данных. Другой обучающий набор данных представляет собой набор из множества заранее заданных обучающих наборов данных, подлежащих использованию при выполнении второго множества итераций обучения. Другой обучающий набор данных выбирается оценщиком-человеком.

[14] В некоторых вариантах осуществления способа на итерации обучения сеть ANN обучается на основе обучающего набора данных, содержащего (а) обучающий запрос, (б) обучающую текстовую входную последовательность и (в) заранее заданную текстовую выходную последовательность. Заранее заданная текстовая выходная последовательность определена оценщиком-человеком. Оценщик-человек определяет заранее заданную текстовую выходную последовательность из контента по меньшей мере одного из обучающих контент-ресурсов, связанных с обучающими сниппетами контента.

[15] В некоторых вариантах осуществления способ дополнительно включает в себя (а) формирование сервером текстового представления запроса на основе звукового представления речевого фрагмента пользователя, указывающего на запрос пользователя, связанного с электронным устройством, (б) определение сервером результатов поиска, соответствующих текстовому представлению запроса, при этом результат поиска содержит сниппет контента и указывает на контент-ресурс, релевантный запросу, (в) формирование сервером текстовой входной последовательности в виде последовательности сниппетов контента на основе сниппетов контента результатов поиска, (г) ввод сервером текстовой входной последовательности и текстового представления запроса в сеть ANN, способную выдавать текстовую выходную последовательность, представляющую собой сводку контента последовательности сниппетов контента и подлежащую использованию в качестве ответа на запрос, (д) формирование сервером звукового представления текстовой выходной последовательности, представляющего собой формируемый компьютером речевой фрагмент ответа, и (е) предоставление сервером электронному устройству, связанному с пользователем, звукового представления текстовой выходной последовательности.

[16] В некоторых вариантах осуществления способа корректировка сети ANN включает в себя применение сервером метода обратного распространения в отношении сети ANN.

[17] В некоторых вариантах осуществления способа он дополнительно включает в себя получение сниппетов контента из соответствующих контент-ресурсов, релевантных запросу.

[18] В некоторых вариантах осуществления способа такое получение включает в себя выполнение поиска с использованием поисковой системы и запроса в качестве входных данных.

[19] В некоторых вариантах осуществления способа такое получение дополнительно включает в себя получение N наиболее релевантных сниппетов, связанных с N лучшими результатами поиска.

[20] В некоторых вариантах осуществления способа он дополнительно включает в себя получение заранее заданной текстовой выходной последовательности.

[21] В некоторых вариантах осуществления способа такое получение осуществляется от оценщика-человека.

[22] В некоторых вариантах осуществления способа он дополнительно включает в себя получение указания на контент-ресурс, использованный оценщиком-человеком для формирования заранее заданной текстовой выходной последовательности.

[23] В некоторых вариантах осуществления способа контент-ресурс используется для проверки заранее заданной текстовой выходной последовательности.

[24] В некоторых вариантах осуществления способа контент-ресурс представляет собой сетевой ресурс, соответствующий обучающему запросу и полученный путем

выполнения поиска с использованием поисковой системы.

[25] Согласно второму аспекту настоящей технологии реализован сервер для обучения сети ANN формированию текстовой выходной последовательности на основе текстовой входной последовательности и запроса. Текстовая входная последовательность формируется в виде последовательности сниппетов контента, полученных из соответствующих контент-ресурсов, релевантных запросу. Текстовая выходная последовательность представляет собой сводку контента последовательности сниппетов контента, которая должна использоваться в качестве ответа на запрос. Сеть ANN реализована на сервере и содержит подсеть кодера и подсеть декодера. Сервер на итерации обучения сети ANN способен вводить обучающий запрос и обучающую текстовую входную последовательность в подсеть кодера. Обучающая текстовая входная последовательность (а) формируется в виде последовательности обучающих сниппетов контента и (б) разделяется на последовательность входных групп. Входная группа связана с обучающим сниппетом контента и содержит слова из обучающего сниппета контента. Сервер на итерации обучения сети ANN способен формировать с использованием подсети кодера кодированное представление обучающей текстовой входной последовательности. Возможность формирования включает в себя возможность формирования сервером выходных данных вида «внимание» для соответствующих слов из обучающей текстовой входной последовательности путем применения маски ограничения внимания в отношении обучающей текстовой входной последовательности. Выходные данные вида «внимание» используются для формирования кодированного представления обучающей текстовой входной последовательности. При формировании выходных данных вида «внимание» для слова из входной группы маска ограничения внимания позволяет «принимать во внимание» только слова из этой входной группы так, чтобы выходные данные вида «внимание» формировались на основе контекста этой входной группы, а не контекстов других входных групп в обучающей текстовой входной последовательности. Сервер на итерации обучения сети ANN способен формировать с использованием подсети декодера декодированное представление обучающей текстовой входной последовательности. Декодированное представление соответствует прогнозируемой текстовой выходной последовательности. Сервер на итерации обучения сети ANN способен формировать оценку штрафа для итерации обучения путем сравнения прогнозируемой текстовой выходной последовательности с заранее заданной текстовой выходной последовательностью. Заранее заданная текстовая выходная последовательность представляет собой заранее заданный ответ на обучающий запрос. Сервер на итерации обучения сети ANN способен корректировать сеть ANN на основе оценки штрафа.

[26] В некоторых вариантах осуществления сервера формируемые выходные данные вида «внимание» дополнительно основываются на контексте из обучающего запроса.

[27] В некоторых вариантах осуществления сервера на итерации обучения сеть ANN обучается на основе обучающего набора данных, содержащего (а) обучающий запрос, (б) обучающую текстовую входную последовательность и (в) заранее заданную текстовую выходную последовательность. Сервер на другой итерации обучения дополнительно способен обучать сеть ANN на основе другого обучающего набора данных. Другой обучающий набор данных содержит (а) обучающий запрос, (б) модифицированную обучающую текстовую входную последовательность и (в) заранее заданную текстовую выходную последовательность. Входные группы в модифицированной обучающей текстовой входной последовательности расположены в порядке, отличном от порядка входных групп в обучающей текстовой входной

последовательности.

[28] В некоторых вариантах осуществления сервера он дополнительно способен формировать обучающий сниппет контента из последовательности обучающих сниппетов контента. Такое формирование включает в себя ввод сервером обучающего запроса и контента из обучающего контент-ресурса в алгоритм MLA, способный выдавать прогнозируемый сниппет контента. Алгоритм MLA обучен формированию прогнозируемого сниппета контента так, чтобы прогнозируемый сниппет контента был подобен заранее заданной текстовой выходной последовательности.

Прогнозируемый сниппет контента представляет собой обучающий сниппет контента.

[29] В некоторых вариантах осуществления сервера на итерации обучения сеть ANN обучается на основе обучающего набора данных, содержащего (а) обучающий запрос, (б) обучающую текстовую входную последовательность и (в) заранее заданную текстовую выходную последовательность. Итерация обучения выполняется на первом этапе обучения сети ANN, включающем в себя первое множество итераций обучения. Сервер дополнительно способен выполнять второй этап обучения сети ANN. Второй этап обучения сети ANN включает в себя второе множество итераций обучения. На другой итерации обучения из второго множества итераций обучения сеть ANN обучается на основе другого обучающего набора данных. Другой обучающий набор данных представляет собой набор из множества заранее заданных обучающих наборов данных, подлежащих использованию при выполнении второго множества итераций обучения. Другой обучающий набор данных выбирается оценщиком-человеком.

[30] В некоторых вариантах осуществления сервера на итерации обучения сеть ANN обучается на основе обучающего набора данных, содержащего (а) обучающий запрос, (б) обучающую текстовую входную последовательность и (в) заранее заданную текстовую выходную последовательность. Заранее заданная текстовая выходная последовательность определена оценщиком-человеком. Оценщик-человек определяет заранее заданную текстовую выходную последовательность из контента по меньшей мере одного из обучающих контент-ресурсов, связанных с обучающими сниппетами контента.

[31] В некоторых вариантах осуществления сервера он дополнительно способен (а) формировать текстовое представление запроса на основе звукового представления речевого фрагмента пользователя, указывающего на запрос пользователя, связанного с электронным устройством, (б) определять результаты поиска, соответствующие текстовому представлению запроса, при этом результат поиска содержит сниппет контента и указывает на контент-ресурс, релевантный запросу, (в) формировать текстовую входную последовательность в виде последовательности сниппетов контента на основе сниппетов контента результатов поиска, (г) вводить текстовую входную последовательность и текстовое представление запроса в сеть ANN, способную выдавать текстовую выходную последовательность, представляющую собой сводку контента последовательности сниппетов контента и подлежащую использованию в качестве ответа на запрос, (д) формировать звуковое представление текстовой выходной последовательности, представляющее собой формируемый компьютером речевой фрагмент ответа, и (е) предоставлять электронному устройству, связанному с пользователем, звуковое представление текстовой выходной последовательности.

[32] В некоторых вариантах осуществления сервера корректировка сети ANN включает в себя применение сервером метода обратного распространения в отношении сети ANN.

[33] В некоторых вариантах осуществления сервера он дополнительно способен

получать сниппеты контента из соответствующих контент-ресурсов, релевантных запросу.

[34] В некоторых вариантах осуществления сервера он с целью такого получения способен выполнять поиск с использованием поисковой системы и запроса в качестве
5 входных данных.

[35] В некоторых вариантах осуществления сервера он с целью такого получения дополнительно способен получать N наиболее релевантных сниппетов, связанных с N лучшими результатами поиска.

[36] В некоторых вариантах осуществления сервера он дополнительно способен
10 получать заранее заданную текстовую выходную последовательность.

[37] В некоторых вариантах осуществления сервера он способен получать заранее заданную текстовую выходную последовательность от оценщика-человека.

[38] В некоторых вариантах осуществления сервера он дополнительно способен
15 получать указание на контент-ресурс, использованный оценщиком-человеком для формирования заранее заданной текстовой выходной последовательности.

[39] В некоторых вариантах осуществления сервера контент-ресурс используется для проверки заранее заданной текстовой выходной последовательности.

[40] В некоторых вариантах осуществления сервера контент-ресурс представляет собой сетевой ресурс, соответствующий обучающему запросу и полученный путем
20 выполнения поиска с использованием поисковой системы.

[41] В контексте настоящего описания термин «сервер» означает компьютерную программу, выполняемую соответствующими аппаратными средствами и способную принимать запросы (например, из устройств) через сеть и выполнять эти запросы или
25 инициировать их выполнение. Соответствующие аппаратные средства могут представлять собой один физический компьютер или одну компьютерную систему, что не существенно для настоящей технологии. В настоящем контексте выражение «сервер» не означает, что каждая задача (например, принятая команда или запрос) или некоторая конкретная задача принимается, выполняется или запускается одним и тем же сервером (т.е. одними и теми же программными и/или аппаратными средствами). Это выражение
30 означает, что любое количество программных средств или аппаратных средств может принимать, отправлять, выполнять или инициировать выполнение любой задачи или запроса либо результатов любых задач или запросов. Все эти программные и аппаратные средства могут представлять собой один сервер или несколько серверов, причем оба эти случая подразумеваются в выражении «по меньшей мере один сервер».

[42] В контексте настоящего описания термин «устройство» означает любое
35 компьютерное аппаратное средство, способное выполнять программы, подходящие для решения поставленной задачи. Таким образом, некоторые (не имеющие ограничительного характера) примеры устройств включают в себя персональные компьютеры (настольные, ноутбуки, нетбуки и т.п.), смартфоны и планшеты, а также
40 сетевое оборудование, такое как маршрутизаторы, коммутаторы и шлюзы. Следует отметить, что в данном контексте устройство, функционирующее как устройство, также может функционировать как сервер для других устройств. Использование термина «устройство» не исключает использования нескольких устройств для приема, отправки, выполнения или инициирования выполнения любой задачи или запроса либо
45 последовательностей любых задач, запросов или шагов любого описанного здесь способа.

[43] В контексте настоящего описания термин «база данных» означает любой структурированный набор данных, независимо от его конкретной структуры,

программного обеспечения для управления базой данных или компьютерных аппаратных средств для хранения этих данных, их применения или обеспечения их использования иным способом. База данных может располагаться в тех же аппаратных средствах, где реализован процесс, обеспечивающий хранение или использование информации, хранящейся в базе данных, либо база данных может располагаться в отдельных аппаратных средствах, таких как специализированный сервер или множество серверов.

[44] В контексте настоящего описания выражение «информация» включает в себя информацию любого рода или вида, допускающую хранение в базе данных. Таким образом, информация включает в себя 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., но не ограничивается ими.

[45] В контексте настоящего описания выражение «компонент» включает в себя обозначение программного обеспечения (подходящего для определенных аппаратных средств), необходимого и достаточного для выполнения определенной функции или нескольких функций.

[46] В контексте настоящего описания выражение «пригодный для использования в компьютере носитель информации» означает носители любого рода и вида, включая оперативное запоминающее устройство (ОЗУ), постоянное запоминающее устройство (ПЗУ), диски (CD-ROM, DVD, гибкие диски, жесткие диски и т.д.), USB-накопители, твердотельные накопители, накопители на магнитных лентах и т.д.

[47] В контексте настоящего описания числительные «первый», «второй», «третий» и т.д. используются лишь для указания различия между существительными, к которым они относятся, но не для описания каких-либо определенных взаимосвязей между этими существительными. Например, должно быть понятно, что использование терминов «первый сервер» и «третий сервер» не подразумевает какого-либо определенного порядка, типа, хронологии, иерархии или классификации, в данном случае, серверов, а также что их использование (само по себе) не подразумевает наличие «второго сервера» в любой ситуации. Кроме того, как встречается в настоящем описании в другом контексте, ссылка на «первый» элемент и «второй» элемент не исключает того, что эти два элемента в действительности могут быть одним и тем же элементом. Таким образом, например, в некоторых случаях «первый» сервер и «второй» сервер могут представлять собой одно и то же программное и/или аппаратное средство, а в других случаях - различные программные и/или аппаратные средства.

[48] Каждый вариант осуществления настоящей технологии относится к по меньшей мере одной из вышеупомянутых целей и/или аспектов, но не обязательно ко всем ним. Должно быть понятно, что некоторые аспекты настоящей технологии, связанные с попыткой достижения вышеупомянутой цели, могут не соответствовать этой цели и/или могут соответствовать другим целям, явным образом здесь не упомянутым.

[49] Дополнительные и/или альтернативные признаки, аспекты и преимущества вариантов осуществления настоящей технологии содержатся в дальнейшем описании, в приложенных чертежах и в формуле изобретения.

Краткое описание чертежей

[50] Дальнейшее описание приведено для лучшего понимания настоящей технологии, а также других аспектов и их признаков, и должно использоваться совместно с приложенными чертежами.

[51] На фиг. 1 представлена схема системы, пригодной для реализации не имеющих

ограничительного характера вариантов осуществления настоящей технологии.

[52] На фиг. 2 представлена страница результатов поисковой системы (SERP, Search Engine Result Page), сформированная сервером системы, представленной на фиг. 1, согласно не имеющим ограничительного характера вариантам осуществления настоящей

5 технологии.

[53] На фиг. 3 представлена структура данных, хранящихся в базе данных системы, представленной на фиг. 1, согласно не имеющим ограничительного характера вариантам осуществления настоящей технологии.

10 [54] На фиг. 4 представлена схема итерации обучения алгоритма MLA, выполняемого системой, представленной на фиг. 1 и способной формировать прогнозируемые сниппеты контента для соответствующих контент-ресурсов, согласно не имеющим ограничительного характера вариантам осуществления настоящей технологии.

15 [55] На фиг. 5 представлена схема итерации обучения сети ANN, выполняемой системой, представленной на фиг. 1 и способной формировать прогнозируемые текстовые выходные последовательности, представляющие собой сводки контента соответствующих контент-ресурсов, согласно не имеющим ограничительного характера вариантам осуществления настоящей технологии.

20 [56] На фиг. 6 представлена блок-схема обработки речевого фрагмента пользователя в системе, представленной на фиг. 1, для предоставления пользователю формируемого компьютером речевого фрагмента, согласно не имеющим ограничительного характера вариантам осуществления настоящей технологии.

[57] На фиг. 7 представлена блок-схема способа обучения сети ANN согласно фиг. 5 в соответствии с другими не имеющими ограничительного характера вариантами осуществления настоящей технологии.

25 **Осуществление изобретения**

[58] На фиг. 1 представлена схема системы 100, пригодной для реализации не имеющих ограничительного характера вариантов осуществления настоящей технологии. Должно быть понятно, что система 100 приведена лишь для демонстрации варианта реализации

30 настоящей технологии. Таким образом, дальнейшее описание системы представляет собой описание примеров, иллюстрирующих настоящую технологию. Это описание не предназначено для определения объема или границ настоящей технологии. В некоторых случаях приводятся полезные примеры модификаций системы 100. Они способствуют пониманию, но также не определяют объема или границ настоящей технологии.

35 [59] Эти модификации не составляют исчерпывающего перечня. Как понятно специалисту в данной области, вероятно, возможны и другие модификации. Кроме того, если в некоторых случаях модификации не описаны (т.е. примеры модификаций отсутствуют), это не означает, что они невозможны и/или что это описание содержит единственно возможный вариант реализации того или иного элемента настоящей

40 технологии. Специалисту в данной области должно быть понятно, что это может быть не так. Кроме того, следует понимать, что система 100 в некоторых случаях может представлять собой упрощенную реализацию настоящей технологии и что такие варианты представлены, чтобы способствовать лучшему ее пониманию. Специалистам в данной области должно быть понятно, что другие варианты осуществления настоящей технологии могут быть значительно сложнее.

45 [60] В общем случае система 100 способна предоставлять ответы на запросы пользователя, которые, можно сказать, в результате обеспечивают «диалог» пользователя и электронного устройства. Например, звуковой сигнал 152 (такой как речевой фрагмент) от пользователя 102 может быть обнаружен электронным

устройством 104 (или просто устройством 104), которое в ответ способно предоставлять звуковые сигналы 154 (такие как речевые фрагменты или формируемые компьютером речевые фрагменты). Таким образом, можно сказать, что в результате этого обеспечивается диалог 150 между пользователем 102 и устройством 104, состоящий из

(а) звукового сигнала 152 и (б) звукового сигнала 154.

[61] Ниже описаны различные элементы системы 100 и то, как эти элементы обеспечивают предоставление звукового сигнала 154 электронным устройством 104 в ответ на звуковой сигнал 152 пользователя 102.

Устройство пользователя

[62] Как упомянуто выше, система 100 содержит устройство 104. На реализацию устройства 104 не накладывается каких-либо особых ограничений. Например, устройство 104 может быть реализовано в виде персонального компьютера (настольного, ноутбука, нетбука и т.д.), беспроводного устройства связи (смартфона, сотового телефона, планшета, умной колонки и т.п.) или сетевого оборудования (маршрутизатора, коммутатора, шлюза). Устройство 104 иногда может называться электронным устройством, оконечным устройством, клиентским электронным устройством или просто устройством. Следует отметить, что связь электронного устройства 104 с пользователем 102 не означает необходимости предлагать или подразумевать какой-либо режим работы, например, вход в систему, регистрацию и т.п.

[63] Предполагается, что устройство 104 содержит аппаратные средства и/или прикладное программное обеспечение и/или встроенное программное обеспечение (либо их сочетание) для (а) обнаружения или фиксации звукового сигнала 152 и (б) предоставления или воспроизведения звукового сигнала 154. Например, устройство 104 может содержать один или несколько микрофонов для обнаружения или фиксации звуковых сигналов 152 и один или несколько громкоговорителей для предоставления или воспроизведения звуковых сигналов 154.

[64] Устройство 104 также содержит аппаратные средства и/или прикладное программное обеспечение и/или встроенное программное обеспечение (либо их сочетание) для выполнения приложения 105 IPA. В общем случае приложение 105 IPA, также известное как чат-бот, предназначено для того, чтобы обеспечивать пользователю 102 возможность отправлять запросы в виде речевых фрагментов (например, в виде звукового сигнала 152) и в ответ предоставлять пользователю 102 ответы в виде речевых фрагментов (например, в виде звукового сигнала 154).

[65] Отправка запросов и предоставление ответов может выполняться приложением 105 IPA с использованием пользовательского интерфейса на естественном языке. В общем случае пользовательский интерфейс на естественном языке приложения 105 IPA может представлять собой человеко-машинный интерфейс любого вида, в котором лингвистические единицы, такие как глаголы, фразы, предложения и т.п., выполняют функции элементов управления пользовательского интерфейса для извлечения, выбора, модификации или иного формирования данных в приложении 105 IPA.

[66] Например, при обнаружении (т.е. фиксации) речевых фрагментов пользователя 102 (например, звукового сигнала 152) устройством 104 приложение 105 IPA может использовать свой пользовательский интерфейс на естественном языке для анализа речевых фрагментов пользователя 102 и извлечения из них данных, указывающих на запросы пользователя. Кроме того, данные, указывающие на полученные устройством 104 ответы, анализируются пользовательским интерфейсом на естественном языке приложения 105 IPA с целью предоставления или воспроизведения речевых фрагментов (например, звукового сигнала 154), указывающих на эти ответы.

[67] Предполагается, что приложение 105 IPA в качестве альтернативы или дополнительно может использовать текстовый человеко-машинный интерфейс, чтобы обеспечивать пользователю 102 возможность отправлять указания на запросы в текстовом виде и в ответ предоставлять пользователю 102 указания на ответы в текстовом виде, без выхода за границы настоящей технологии.

Сеть связи

[68] В представленном для иллюстрации не имеющем ограничительного характера примере системы 100 устройство 104 связано с сетью 110 связи для доступа и передачи пакетов данных серверу 106 и/или другим веб-ресурсам (не показаны) или от них. В некоторых не имеющих ограничительного характера вариантах осуществления настоящей технологии в качестве сети 110 связи может использоваться сеть Интернет. В других не имеющих ограничительного характера вариантах осуществления данной технологии сеть 110 связи может быть реализована иначе, например, в виде любой глобальной сети связи, локальной сети связи, частной сети связи и т.п. Реализация линии связи (не обозначена) между устройством 104 и сетью 110 связи зависит, среди прочего, от реализации устройства 104.

[69] Лишь в качестве не имеющего ограничительного характера примера, в тех вариантах осуществления настоящей технологии, в которых устройство 104 реализовано как беспроводное устройство связи (такое как смартфон), линия связи может быть реализована как беспроводная линия связи (такая как канал сети связи 3G, канал сети связи 4G, Wireless Fidelity или сокращенно WiFi[®], Bluetooth[®] и т.п.). В тех примерах, где устройство 104 реализовано как ноутбук, линия связи может быть беспроводной (такой как Wireless Fidelity или сокращенно WiFi[®], Bluetooth[®] и т.д.) или проводной (такой как соединение на основе Ethernet).

Множество серверов ресурсов

[70] Как описано выше, множество серверов 108 ресурсов может быть доступно через сеть 110 связи. Множество серверов 108 ресурсов может быть реализовано в виде традиционных компьютерных серверов. В не имеющем ограничительного характера примере осуществления настоящей технологии сервер из множества серверов 108 ресурсов может быть реализован в виде сервера Dell[™] PowerEdge[™], работающего под управлением операционной системы Microsoft[™] Windows Server[™]. Сервер из множества серверов 108 ресурсов также может быть реализован с использованием любых других подходящих аппаратных средств и/или прикладного программного обеспечения и/или встроенного программного обеспечения либо их сочетания.

[71] Множество серверов 108 ресурсов способно содержать ресурсы (или веб-ресурсы), которые могут быть доступны устройству 104 и/или серверу 106. На вид ресурсов, содержащихся на множестве серверов 108 ресурсов, не накладывается каких-либо ограничений. Тем не менее, в некоторых вариантах осуществления настоящей технологии эти ресурсы могут содержать цифровые документы или просто документы, представляющие собой веб-страницы.

[72] Например, множество серверов 108 ресурсов может содержать веб-страницы, т.е. множество серверов 108 ресурсов может хранить документы, которые представляют собой веб-страницы и доступны устройству 104 и/или серверу 106. Документ может быть составлен на языке разметки и может, среди прочего, содержать (а) контент соответствующей веб-страницы и (б) машиночитаемые команды для отображения соответствующей веб-страницы (содержащегося на ней контента). Таким образом, в по меньшей мере некоторых вариантах осуществления настоящей технологии можно сказать, что множество серверов 108 ресурсов может содержать множество контент-

ресурсов, каждый из которых связан с соответствующим контентом.

[73] Устройство 104 может обращаться к серверу из множества серверов 108 ресурсов с целью получения документа (например, контент-ресурса), хранящегося на этом сервере. Например, пользователь 102 может ввести веб-адрес, связанный с веб-страницей, в браузерном приложении (не показано) устройства 104. В ответ устройство 104 может обратиться к серверу ресурсов, содержащему эту веб-страницу, с целью получения документа, представляющего собой эту веб-страницу, для отображения контента этой веб-страницы с использованием браузерного приложения.

[74] Следует отметить, что сервер 106 также может обращаться к серверу из множества серверов 108 ресурсов с целью получения документа (например, контент-ресурса), хранящегося на этом сервере ресурсов. Назначение сервера 106, осуществляющего доступ и получение документов от множества серверов 108 ресурсов, более подробно описано ниже.

Сервер

[75] Как упомянуто выше, система 100 также содержит сервер 106, который может быть реализован в виде традиционного компьютерного сервера. В примере осуществления настоящей технологии сервер 106 может быть реализован в виде сервера Dell™ PowerEdge™, работающего под управлением операционной системы Microsoft™ Windows Server™. Должно быть понятно, что сервер 106 может быть реализован с использованием любых других подходящих аппаратных средств, прикладного программного обеспечения и/или встроенного программного обеспечения, либо их сочетания. В представленных не имеющих ограничительного характера вариантах осуществления настоящей технологии сервер 106 представляет собой один сервер. В других не имеющих ограничительного характера вариантах осуществления настоящей технологии функции сервера 106 могут быть распределены между несколькими серверами.

[76] В общем случае сервер 106 способен (а) получать от устройства 104 данные, связанные со звуковым сигналом 152, (б) анализировать эти данные и в ответ (в) определять данные для формирования звукового сигнала 154. Например, сервер 106 может получать данные, указывающие на запрос пользователя, и в ответ может формировать данные, указывающие на ответ на этот запрос пользователя. С этой целью сервер 106, среди прочего, способен выполнять одну или несколько компьютерных процедур, которые в этом документе называются поисковой системой 130.

[77] В некоторых вариантах осуществления изобретения сервер 106 может управляться и/или администрироваться поставщиком услуг поисковой системы (не показан), таким как оператор поисковой системы Yandex™. Сервер 106 может содержать поисковую систему 130 для выполнения одного или нескольких поисков в ответ на запросы, отправленные пользователями поисковой системы 130.

[78] В некоторых вариантах осуществления изобретения сервер 106 может получать данные от устройства 104, указывающие на запрос пользователя 102. Например, приложение 105 IPA устройства 104 может формировать цифровое звуковое представление для звукового сигнала 152 (например, для речевого фрагмента пользователя) и отправлять это цифровое звуковое представление через сеть 110 связи серверу 106. В этом примере сервер 106 может формировать текстовое представление запроса, отправленного пользователем 102, на основе цифрового звукового представления речевого фрагмента пользователя и выполнять поиск на основе сформированного таким образом текстового представления запроса. В по меньшей мере некоторых вариантах осуществления настоящей технологии пользователь 102

может использовать электронное устройство 104 и вручную вводить текстовое представление запроса, а ответ, предоставляемый пользователю 102, может быть предоставлен в виде цифрового звукового представления без выхода за границы настоящей технологии.

5 [79] Следует отметить, что преобразование текста в речь для цифрового звукового представления речевого фрагмента пользователя может быть выполнено локально в устройстве 104 (например, приложением 105 IPA). Следовательно, устройство 104 дополнительно или в качестве альтернативы может отправлять текстовое представление запроса серверу 106 в дополнение к цифровому звуковому представлению или вместо
10 него.

[80] Как описано выше, сервер 106 может использовать текстовое представление запроса для выполнения поиска (соответствующего запросу) и формировать таким образом результаты поиска, релевантные запросу. Например, сервер 106, использующий поисковую систему 130, может формировать данные, указывающие на один или
15 несколько результатов поиска, релевантных запросу пользователя 102.

[81] Для лучшей иллюстрации этого дальнейшее описание приведено со ссылкой на фиг. 2, где содержится представление 200 страницы SERP. Когда пользователь 102 отправляет свой запрос поисковой системе 130, например, с помощью браузерного приложения, поисковая система 130 обычно используется для формирования так
20 называемой страницы SERP, содержащей один или несколько результатов поиска, релевантных запросу. Как показано на фиг. 2, поисковая система 130 способна использовать текстовое представление 202 запроса для формирования ранжированного списка результатов поиска, который в данном случае содержит первый результат 210 поиска, второй результат 220 поиска и третий результат 230 поиска.

25 [82] Следует отметить, что в представленном на фиг. 2 не имеющем ограничительного характера примере (а) первый результат 210 поиска связан с первым контент-ресурсом 212 и с первым сниппетом 214 контента, (б) второй результат 220 поиска связан со вторым контент-ресурсом 222 и содержит второй сниппет 224 контента, (в) третий результат 230 поиска связан с третьим контент-ресурсом 232 и с третьим сниппетом
30 234 контента.

[83] Поисковая система 130 способна определять контент-ресурсы, релевантные запросу пользователя, и ранжировать эти контент-ресурсы на основе, среди прочего, релевантности их контента запросу, как это известно в данной области техники.

[84] В по меньшей мере некоторых не имеющих ограничительного характера
35 вариантах осуществления настоящей технологии разработчики настоящей технологии разработали способы и системы для формирования ориентированных на пользователя сниппетов контента для соответствующих контент-ресурсов. Разработчики настоящей технологии установили, что предоставление пользователю более лаконичных сниппетов контента (т.е. кратких и содержательных) может повысить удовлетворенность
40 пользователя от сервисов поисковой системы. Такие ориентированные на пользователя сниппеты контента позволяют уменьшить интервал времени, требуемый для анализа результатов поиска, предоставленных поисковой системой.

[85] С этой целью в по меньшей мере некоторых не имеющих ограничительного характера вариантах осуществления настоящей технологии сервер 106 способен
45 выполнять одну или несколько компьютерных процедур, которые ниже называются механизмом 135 сниппетов. В общем случае механизм 135 сниппетов способен формировать для контент-ресурса соответствующий сниппет контента, представляющий собой краткую сводку контента этого ресурса.

[86] Разработчики настоящей технологии также установили, что механизм 135
 5 снippetов, способный формировать краткие снippetы контента для различных контент-
 ресурсов, может быть использован в контексте обработки речевых запросов. Очевидно,
 что в некоторых вариантах осуществления настоящей технологии, где устройство 104
 10 представляет собой, например, умную колонку, желательно предоставлять пользователю
 102 голосовые сервисы поисковой системы. В этих вариантах осуществления изобретения
 формирование лаконичных снippetов контента может быть использовано для создания
 формируемых компьютером речевых фрагментов, которые более кратко и
 содержательно передают контент, релевантный запросу. Ниже более подробно описано
 10 формирование механизмом 135 снippetов кратких снippetов контента для
 соответствующих контент-ресурсов.

[87] Также следует отметить, что разработчики настоящей технологии установили,
 что в контексте речевых запросов для дальнейшего повышения удовлетворенности
 пользователя от краткости и содержательности формируемых компьютером речевых
 15 фрагментов может потребоваться дополнительная обработка таких лаконичных
 снippetов контента. Понятно, что предоставляемый в звуковом формате
 ранжированный список снippetов контента менее удобен, чем визуально отображаемая
 пользователю страница SERP.

[88] Исходя из этого, разработчики настоящей технологии разработали способы и
 20 системы, позволяющие серверу 106, в известном смысле, обобщать контент множества
 контент-ресурсов, релевантных запросу пользователя, и предоставлять эту сводку
 контента пользователю 102 в звуковом формате (т.е. предоставлять формируемый
 компьютером речевой фрагмент). В по меньшей мере некоторых не имеющих
 ограничительного характера вариантах осуществления настоящей технологии
 25 разработчики настоящей технологии разработали сеть 140 ANN, способную
 формировать текстовую выходную последовательность, представляющую собой такую
 сводку контента.

[89] Также предполагается, что по меньшей мере некоторые описанные здесь способы
 и системы способны использовать лаконичные снippetы контента (сформированные
 30 механизмом 135 снippetов) для обучения сети 140 ANN формированию краткой и
 содержательной сводки контента для множества контент-ресурсов, релевантных запросу.
 Ниже более подробно описано формирование сетью 140 ANN этой текстовой выходной
 последовательности, представляющей собой сводку контента множества контент-
 ресурсов.

[90] Следует отметить, что сервер 106 способен предоставлять устройству 104 данные,
 указывающие на сводку контента в текстовом формате, которая должна обрабатываться
 35 приложением 105 IPA с целью формирования соответствующего цифрового звукового
 представления (например, звукового сигнала 154) и воспроизводиться пользователю
 102. Тем не менее, в других вариантах осуществления изобретения сервер 106 может
 40 формировать цифровое звуковое представление сводки контента на основе
 соответствующей текстовой выходной последовательности и таким образом
 предоставлять данные, указывающие на сводку контента в цифровом звуковом формате.

Система базы данных

[91] Сервер 106 связан с системой 120 базы данных. В общем случае система 120 базы
 45 данных может хранить информацию, полученную и/или сформированную сервером
 106 во время обработки. Например, система 120 базы данных способна принимать от
 сервера 106 данные, полученные и/или сформированные сервером 106 во время
 обработки, для временного и/или постоянного хранения и выдавать сохраненные

данные серверу 106 для дальнейшего их использования. Для лучшей иллюстрации этого дальнейшее описание ссылается на фиг. 3, где приведено представление 300 данных, хранящихся в системе 120 базы данных, согласно по меньшей мере некоторым не имеющим ограничительного характера вариантам осуществления настоящей технологии.

[92] Система 120 базы данных способна хранить данные 310 поисковой системы. В общем случае данные 310 поисковой системы, хранящиеся в системе 120 базы данных, содержат данные, позволяющие серверу 106 предоставлять сервисы поисковой системы пользователям поисковой системы 130. В некоторых вариантах осуществления изобретения данные 310 поисковой системы могут содержать данные о большом количестве контент-ресурсов (например, документов), полученных приложением обходчика (не показано). Таким образом, система 120 базы данных может предоставлять серверу 106 доступ к этому большому количеству документов во время поиска документов.

[93] Следует отметить, что сервер 106 может выполнять приложение обходчика в качестве части поисковой системы 130. В общем случае приложение обходчика может использоваться сервером 106 для «посещения» ресурсов, доступных через сеть 110 связи, и для получения или загрузки документов с целью их дальнейшего использования. Например, приложение обходчика может быть использовано сервером 106 для доступа к множеству серверов 108 ресурсов и для получения или загрузки документов, представляющих собой веб-страницы, содержащиеся на множестве серверов 108 ресурсов. Предполагается, что приложение обходчика может периодически выполняться сервером 106 с целью получения или загрузки документов, которые были обновлены и/или стали доступными через сеть 110 связи после предыдущего выполнения приложения обходчика.

[94] Следует отметить, что данные о большом количестве контент-ресурсов могут храниться в системе 120 базы данных в индексированном формате, т.е. система 120 базы данных может хранить известную в данной области техники структуру индексации, такую как инвертированный индекс контент-ресурсов. Например, система 120 базы данных может хранить множество списков вхождений, связанных с соответствующими терминами (например, со словами), идентифицированными в контенте контент-ресурсов. По сути, список вхождений указывает на множество контент-ресурсов, в контенте которых содержится соответствующий термин.

[95] В некоторых вариантах осуществления изобретения данные 310 поисковой системы, хранящиеся в системе 120 базы данных, могут содержать информацию о поисках, ранее выполненных поисковой системой 130. Например, данные 310 поисковой системы могут содержать данные запроса, связанные с соответствующими запросами. Данные запроса, связанные с запросом, могут быть различных видов и на них не накладывается каких-либо ограничений. Например, данные запроса могут содержать (не ограничиваясь этим):

- популярность запроса;
- частоту отправки запроса;
- количество «кликов», связанных с запросом;
- указания на другие отправленные запросы, связанные с запросом;
- указания на документы, связанные с запросом;
- другие статистические данные, связанные с запросом;
- текстовое представление запроса;
- количество символов в запросе;
- звуковое представление (или представления) запроса;

- другие присущие запросу характеристики.

[96] Также предполагается, что данные 310 поисковой системы, хранящиеся в системе 120 базы данных, могут содержать информацию о контент-ресурсах, предоставленных поисковой системой 130 в качестве результатов поиска. Данные ресурса, связанные с контент-ресурсом, могут быть различных видов и на них не накладывается каких-либо ограничений. Например, данные ресурса могут содержать (не ограничиваясь этим):

- популярность контент-ресурса;
- коэффициент «кликов» для контент-ресурса;
- время на «клик», связанное с контент-ресурсом;
- указания на другие документы, связанные с контент-ресурсом;
- указания на запросы, связанные с контент-ресурсом;
- другие статистические данные, связанные с контент-ресурсом;
- текстовые данные, связанные с контент-ресурсом;
- другие присущие документу характеристики контент-ресурса.

[97] Система 120 базы данных может хранить данные 330 оценщика-человека. В общем случае данные 330 оценщика-человека содержат информацию, по меньшей мере частично сформированную оценщиками-людьми и используемую сервером 106 для обучения алгоритма 400 MLA (см. фиг. 4), реализованного в виде части механизма 135 сниппетов, и сети 140 ANN. Ниже более подробно описано использование сервером 106 данных 330 оценщика-человека для обучения алгоритма 400 MLA механизма 135 сниппетов и сети 140 ANN сервера 106.

[98] Следует отметить, что данные 330 оценщика-человека могут содержать множество оцененных людьми наборов данных. Оцененный человеком набор данных может содержать (а) указание на запрос (Q), (б) указание на контент-ресурс, релевантный этому запросу, и (в) указание на оцененный человеком сниппет контента для соответствующего контент-ресурса. Оцененный человеком набор данных может быть сформирован путем предоставления оценщику-человеку запроса и последующего сбора данных, предоставленных оценщиком-человеком. Следует отметить, что данные 330 оценщика-человека могут содержать большое количество оцененных людьми наборов данных для обучения, как описано ниже.

[99] Например, в ответ на запрос оценщик-человек может использовать поисковую систему 130 для выполнения поиска, соответствующего запросу, и получения множества результатов поиска, связанных с соответствующими контент-ресурсами, которые поисковая система 130 определила как релевантные запросу. Затем оценщику-человеку ставится задача анализа контента предоставленных в качестве результатов поиска контент-ресурсов и формирования оцененного человеком сниппета контента для этого запроса на основе контента по меньшей мере одного из контент-ресурсов, предоставленных в качестве результатов поиска. В некоторых не имеющих ограничительного характера вариантах осуществления настоящей технологии оценщику-человеку также ставится задача предоставить указание на по меньшей мере один из предоставленных в качестве результатов поиска контент-ресурсов, использованный для формирования сводки (например, для проверки, контроля качества и т.п.).

[100] Для лучшей иллюстрации этого можно предположить, что оценщик-человек использует контент второго по релевантности этому запросу контент-ресурса для запроса с целью формирования оцененного человеком сниппета контента. В этом случае оцененный человеком набор данных может содержать (а) указание на запрос, (б) указание на второй по релевантности контент-ресурс и (в) указание на оцененный человеком сниппет контента, сформированный на основе контента второго по

релевантности контент-ресурса.

[101] В некоторых не имеющих ограничительного характера вариантах осуществления изобретения оцененный человеком сниппет контента может представлять собой оригинальную текстовую последовательность, обнаруженную оценщиком-человеком в контенте соответствующего контент-ресурса. Например, если запрос представляет собой выражение «John Malkovich», то оценщик-человек может определить в качестве оцененного человеком сниппета контента следующую текстовую последовательность из второго по релевантности контент-ресурса: «John Gavin Malkovich was born on December 9, 1953 in the United States of America. He is a voice actor, director, producer, and fashion designer. He received Academy award nominations for his performances in Places in the Heart (1984) and In the Line of Fire (1993)».

[102] В других не имеющих ограничительного характера вариантах осуществления изобретения оцененный человеком сниппет контента может представлять собой текстовую последовательность на основе контента соответствующего контент-ресурса, частично модифицированную оценщиком-человеком для краткости. Например, если запрос представляет собой выражение «John Malkovich», то оценщик-человек может определить следующую текстовую последовательность из второго по релевантности контент-ресурса: «John Gavin Malkovich was born on December 9, 1953 in the United States of America. He is a voice actor, director, producer, and fashion designer. He received Academy award nominations for his performances in Places in the Heart (1984) and In the Line of Fire (1993)» и затем может частично модифицировать эту текстовую последовательность для формирования на ее основе более лаконичного оцененного человеком сниппета контента. Например, в этом случае оценщик-человек может сформировать следующий оцененный человеком сниппет контента: «John Gavin Malkovich is an Academy award nominated American actor born on December 9, 1953, and who is also a director, producer, and fashion designer». Это означает, что в некоторых случаях оценщик-человек может частично модифицировать оригинальную текстовую последовательность, обнаруженную в контенте соответствующего контент-ресурса, таким образом, чтобы она представляла собой сводку контента, более краткую и содержательную (т.е. более лаконичную), чем оригинальная текстовая последовательность из соответствующего контент-ресурса.

[103] Ниже более подробно описано формирование механизмом 135 сниппетов, содержащим алгоритм 400 MLA, кратких сниппетов контента для соответствующих контент-ресурсов.

Алгоритм MLA механизма сниппетов

[104] Как описано выше и показано на фиг. 4, механизм 135 сниппетов содержит алгоритм 400 MLA. В общем случае алгоритмы MLA способны обучаться на обучающих выборках и осуществлять прогнозирование на основе новых (неизвестных) данных. Алгоритмы MLA обычно используются для первоначального построения модели на основе обучающих входных данных, чтобы затем на основе данных осуществлять прогнозы или принимать решения, выраженные в виде выходных данных, а не выполнять статические машиночитаемые команды. Алгоритмы MLA могут использоваться в качестве моделей оценивания, моделей ранжирования, моделей классификации и т.п.

[105] Должно быть понятно, что в некоторых не имеющих ограничительного характера вариантах осуществления настоящей технологии для реализации алгоритма 400 MLA могут использоваться алгоритмы MLA различных видов с различными архитектурами и/или топологиями. Тем не менее, в контексте настоящей технологии можно сказать, что алгоритм 400 MLA представляет собой алгоритм MLA, который

на основе пары запрос-ресурс обучается формированию прогнозируемого сниппета контента (например, в текстовом формате).

[106] В общем случае реализация алгоритма 400 MLA сервером 106 может быть разделена на два основных этапа: этап обучения и этап использования. Сначала алгоритм 400 MLA обучается на этапе обучения. Затем, когда алгоритму 400 MLA известно, какие предполагаются входные данные и какие должны выдаваться выходные данные, алгоритм 400 MLA выполняется с реальными данными на этапе использования. Как описано ниже, предполагается, что сервер 106 также может использовать алгоритм 400 MLA для формирования по меньшей мере некоторых обучающих данных для обучения сети 140 ANN.

[107] На фиг. 4 приведено представление 480 одной итерации обучения алгоритма 400 MLA. Тем не менее, следует отметить, что сервер 106 может выполнять большое количество итераций обучения на основе соответствующих обучающих наборов данных подобно тому, как сервер 106 способен выполнять одну итерацию обучения, представленную на фиг. 4.

[108] Во время представленной итерации обучения сервер 106 способен обучать алгоритм 400 MLA на основе обучающего набора 410 данных. В по меньшей мере некоторых не имеющих ограничительного характера вариантах осуществления настоящей технологии обучающий набор 410 данных может соответствовать соответствующему оцененному человеком набору данных из множества оцененных людьми наборов данных, хранящихся в системе 120 базы данных. Иными словами, обучающий набор 410 данных может содержать данные, указывающие (а) на запрос 416, (б) на соответствующий контент-ресурс 414 и (в) на соответствующий целевой сниппет 412 контента (например, оцененный человеком сниппет контента для соответствующей пары запрос-ресурс).

[109] Сервер 106 способен использовать (а) запрос 416 и соответствующий контент-ресурс 414 в качестве входных данных алгоритма 400 MLA и (б) соответствующий целевой сниппет 412 контента в качестве обучающей цели. В некоторых вариантах осуществления изобретения предполагается, что данные, указывающие на запрос 416 и на соответствующий контент-ресурс 414, могут быть введены в алгоритм 400 MLA в векторном формате. Например, контент запроса 416 и соответствующего контент-ресурса 414 может быть введен как векторные представления на основе соответствующих представлений вида «мешок слов» (bag-of-words).

[110] В ответ на введенные данные, представляющие запрос 416 и соответствующий контент-ресурс 414, алгоритм 400 MLA способен выдавать прогнозируемый сниппет 420 контента. Иными словами, прогнозируемый сниппет 420 контента представляет собой сформированный алгоритмом 400 MLA прогноз сниппета контента для соответствующего контент-ресурса 414, например, когда контент-ресурс 414 предоставляется в качестве результата поиска, релевантного запросу 416. Предполагается, что прогнозируемый сниппет 420 контента может выдаваться алгоритмом 400 MLA в векторном формате без выхода за границы настоящей технологии.

[111] Затем сервер 106 способен формировать оценку 430 штрафа для итерации обучения на основе прогнозируемого сниппета 420 контента и целевого сниппета 412 контента (т.е. на основе сравнения прогноза с целью). На вид метрики штрафа, используемой для формирования оценки 430 штрафа на основе прогнозируемого сниппета 420 контента и целевого сниппета 412 контента, не накладывается каких-либо особых ограничений. Тем не менее, следует отметить, что оценка 430 штрафа указывает

на различие между целевым сниппетом 412 контента и прогнозируемым сниппетом 420 контента. Также можно сказать, что оценка 430 штрафа может указывать на степень сходства или различия целевого сниппета 412 контента и прогнозируемого сниппета 420 контента.

5 [112] На итерации обучения сервер 106 способен обучать алгоритм 400 MLA на основе оценки 430 штрафа. Например, сервер 106 может использовать метод обратного распространения для корректировки алгоритма 400 MLA на основе оценки 430 штрафа. Таким образом, можно сказать, что в результате выполнения итерации обучения сервером 106 алгоритм 400 MLA обучается формированию соответствующих
10 прогнозируемых сниппетов контента, схожих с соответствующими целевыми сниппетами контента (т.е. формированию прогнозов, более близких к соответствующим целям). С учетом того, что целевые сниппеты контента могут представлять собой соответствующие оцененные людьми сниппеты контента из системы 120 базы данных, алгоритм 400 MLA таким образом может быть обучен формированию прогнозируемых сниппетов контента,
15 схожих с оцененными людьми сниппетами контента (например, с краткими сниппетами контента).

[113] Как описано выше, выполнив большое количество итераций обучения, сервер 106 может применять обученный алгоритм 400 MLA на этапе его использования. В не имеющем ограничительного характера примере сервер 106 может использовать алгоритм
20 400 MLA при формировании сниппетов контента для результатов поиска, соответствующих запросу, отправленному поисковой системе 130. Для лучшей иллюстрации этого на фиг. 2 показано, что сервер 106 может использовать обученный таким образом алгоритм 400 MLA, чтобы формировать:

- первый сниппет 214 контента для первого результата 210 поиска на основе запроса
25 202 и первого контент-ресурса 212;
- второй сниппет 224 контента для второго результата 220 поиска на основе запроса 202 и второго контент-ресурса 222;
- третий сниппет 234 контента для третьих результатов 230 поиска на основе запроса 202 и третьего контент-ресурса 232.

30 [114] Кроме того, в по меньшей мере некоторых не имеющих ограничительного характера вариантах осуществления настоящей технологии в дополнение к использованию алгоритма 400 MLA при формировании страницы SERP поисковой системой 130 сервера 106, алгоритм 400 MLA также может использоваться сервером 106 для обучения сети 140 ANN. Ниже более подробно описано обучение и последующее
35 использование сети 140 ANN сервером 106 для формирования сводки контента страницы SERP.

Сети ANN

[115] Как описано выше, сервер 106 может содержать сеть 140 ANN. В общем случае сети NN представляют собой особый класс алгоритмов MLA и состоят из
40 взаимосвязанных групп искусственных «нейронов», обрабатывающих информацию с использованием коннекционного подхода к вычислениям. Сети NN используются для моделирования сложных взаимосвязей между входными и выходными данными (без фактической информации об этих взаимосвязях) или для поиска закономерностей в данных. Сети NN сначала адаптируются на этапе обучения, когда они обеспечиваются
45 известным набором входных данных и информацией для адаптации сети NN с целью формирования соответствующих выходных данных (в ситуации, для которой выполняется попытка моделирования). На этом этапе обучения сеть NN адаптируется к изучаемой ситуации и изменяет свою структуру так, чтобы сеть NN была способна

обеспечивать адекватное предсказание выходных данных для входных данных в новой ситуации (на основе того, что было выучено). Таким образом, вместо попытки определения сложных статистических распределений или математических алгоритмов для данной ситуации, сеть NN пытается предоставить «интуитивный» ответ на основе «восприятия» ситуации. Таким образом, сеть NN представляет собой своего рода обученный черный ящик, который может быть использован в ситуациях, когда содержимое ящика может быть менее важным, чем то, что этот ящик предоставляет адекватные ответы для входных данных. Например, сети NN широко используются для оптимизации распределения веб-трафика между серверами и при обработке данных, включая фильтрацию, кластеризацию, разделение сигналов, сжатие, векторизацию и т.п.

[116] Следует отметить, что модели вида «кодер-декодер» представляют собой класс моделей на основе сетей NN. Модели вида «кодер-декодер» обычно состоят из (а) подсети кодера, используемой для кодирования входных данных в их «скрытые представления», и (б) подсети декодера, используемой для декодирования этих скрытых представлений в выходные данные. Они часто используются для решения задач, которые в данной области техники называются задачами прогнозирования «из последовательности в последовательность» (seq2seq). Такие задачи включают в себя машинный перевод, реферирование текстов, реализацию чатботов и т.п.

[117] На практике некоторые сети NN также называются моделями seq2seq и способны преобразовывать входную последовательность данных в выходную последовательность данных. В моделях seq2seq могут использоваться рекуррентные нейронные сети (RNN, Recurrent NN), например, для того, чтобы, в известном смысле, «извлекать» контекст, в котором входные данные используются в соответствующей последовательности. Для простоты можно сказать, что подсеть кодера модели seq2seq преобразует каждые входные данные в соответствующий скрытый вектор (содержащий входные данные и контекст), а подсеть декодера выполняет обратный процесс и преобразует вектор в выходные данные (с использованием предыдущих выходных данных в качестве входного контекста).

[118] Модели seq2seq более современного класса называются трансформерами (transformer). Подобно моделям на основе сетей RNN, трансформеры предназначены для обработки последовательных данных в таких задачах, как перевод и реферирование текстов. Тем не менее, в отличие от моделей на основе сетей RNN, трансформеры не требуют, чтобы эти последовательные данные обрабатывались в конкретном порядке. Например, если входные данные представляют собой предложение на естественном языке, то трансформеру не требуется обрабатывать его начало до обработки конца. Благодаря этой особенности трансформеры допускают гораздо большее распараллеливание, чем модели на основе сетей RNN и, следовательно, позволяют уменьшить время обучения. С момента разработки трансформеров они использовались для решения различных задач, связанных с обработкой естественного языка (NLP, Natural Language Processing). В качестве примера такого варианта реализации можно привести модель на основе трансформера, используемую в системе «Представления двунаправленного кодера из трансформеров» (BERT, Bidirectional Encoder Representations from Transformers), которая обучена на больших языковых наборах данных и может быть точно настроена для конкретных языковых задач.

[119] В контексте настоящей технологии сеть 140 ANN может быть реализована в виде модели «кодер-декодер», содержащей подсеть кодера и подсеть декодера.

[120] Предполагается, что подсеть кодера сети 140 ANN способна реализовывать

конкретный механизм внимания. Например, подсеть кодера может содержать слой внимания, способный применять маску ограничения внимания в отношении входных данных. Как описано ниже, при определении контекста для слова из входной последовательности маска ограничения внимания может «принимать во внимание» слова из входной группы, к которой относится слово из входной последовательности, и поэтому позволяет учитывать только контекст соответствующей входной группы, а не другие контексты других входных групп входной последовательности. Ниже со ссылкой на фиг. 5 более подробно описаны обучение сети 140 ANN, определение входных групп из входной последовательности и настройка маски ограничения внимания подсети кодера.

Обучение сети ANN

[121] На фиг. 5 представлена одна итерация обучения сети 140 ANN. Тем не менее, следует отметить, что сервер 106 может выполнять большое количество итераций обучения сети 140 ANN. Это большое количество итераций обучения сети 140 ANN может быть выполнено на основном этапе обучения сети 140 ANN. Предполагается, что в по меньшей мере некоторых не имеющих ограничительного характера вариантах осуществления настоящей технологии сеть 140 ANN может быть обучена на нескольких этапах обучения или на этапе обучения, содержащем несколько стадий.

[122] Предполагается, что сеть 140 ANN может выполнять соответствующее множество итераций обучения во время каждого этапа обучения. В некоторых не имеющих ограничительного характера вариантах осуществления настоящей технологии сеть 140 ANN может иметь (а) предварительный этап обучения, во время которого для настройки модели используется предварительный набор обучающих наборов данных, (б) основной этап обучения, во время которого для обучения модели формированию удовлетворительных прогнозов используется основной набор обучающих наборов данных, и (в) этап обучения для предотвращения «галлюцинаций», во время которого заранее заданные обучающие наборы данных используются для такой корректировки модели, чтобы модель избегала формирования «необычных» прогнозов. Следует отметить, что итерация обучения во время любого этапа обучения сети 140 ANN может быть выполнена подобно тому, как выполняется итерация обучения, представленная на фиг. 5. Тем не менее, как описано ниже, итерации обучения каждого соответствующего этапа обучения сети 140 ANN могут быть выполнены сервером 106 на соответствующих обучающих наборах данных для соответствующих целей обучения.

[123] Сеть 140 ANN содержит подсеть 602 кодера и подсеть 604 декодера. На итерации обучения сервер 106 может вводить в подсеть 602 кодера обучающий запрос 642 и обучающую текстовую входную последовательность 610.

[124] Следует отметить, что обучающая текстовая входная последовательность 610 формируется в виде последовательности обучающих сниппетов контента, включая i -й обучающий сниппет 612 контента, первый обучающий сниппет 614 контента, второй обучающий сниппет 616 контента, третий обучающий сниппет 618 контента и j -й обучающий сниппет 620 контента. В одном не имеющем ограничительного характера варианте осуществления настоящей технологии обучающая текстовая входная последовательность 610 может формироваться в виде последовательности пятнадцати обучающих сниппетов контента. Тем не менее, обучающая текстовая входная последовательность из обучающего набора данных сети 140 ANN может содержать больше или меньше пятнадцати обучающих сниппетов контента без выхода за границы настоящей технологии.

[125] Следует отметить, что обучающие сниппеты контента из обучающей текстовой

входной последовательности 610 могут формироваться сервером 106 разными способами. В одном не имеющем ограничительного характера примере сервер 106 может использовать поисковую систему 130 для формирования обучающих сниппетов контента для набора результатов поиска, соответствующих обучающему запросу 642.

5 В некоторых вариантах осуществления настоящей технологии сервер 106 может формировать обучающие сниппеты контента для набора результатов поиска согласно известным способам.

[126] В других вариантах осуществления настоящей технологии соответствующие обучающие сниппеты контента из обучающей текстовой входной последовательности 610 могут формироваться алгоритмом 400 MLA механизма 135 сниппетов. Например, сервер 106 может определять пятнадцать контент-ресурсов (или любое другое подходящее количество контент-ресурсов), релевантных обучающему запросу 642, и с использованием алгоритма 400 MLA формировать соответствующий сниппет контента для каждого контент-ресурса из числа пятнадцати контент-ресурсов, как описано выше.

15 Затем сервер 106 может формировать обучающую текстовую входную последовательность 610 путем формирования последовательности сниппетов контента соответствующих контент-ресурсов (путем конкатенации). Таким образом, как описано выше, можно сказать, что сервер 106 может использовать алгоритм 400 MLA для формирования по меньшей мере некоторых обучающих данных для обучения сети 140 ANN без выхода за границы настоящей технологии.

[127] Также следует отметить, что обучающая текстовая входная последовательность 610 разделяется на последовательность входных групп. В контексте настоящей технологии входная группа связана с обучающим сниппетом контента и содержит слова из обучающего сниппета контента. В одном не имеющем ограничительного

25 характера примере сервер 106 может определять слова из обучающей текстовой входной последовательности 610, подлежащие объединению во входную группу. С этой целью сервер 106 может использовать соответствующие обучающие сниппеты контента из обучающей текстовой входной последовательности 610 для сопоставления слов из обучающей текстовой входной последовательности 610 с соответствующими входными группами. В представленном примере последовательность входных групп содержит:

- i -ю входную группу 632, содержащую слова из i -го обучающего сниппета 612 контента;
- первую входную группу 634, содержащую слова из первого обучающего сниппета 614 контента;
- 35 - вторую входную группу 636, содержащую слова из второго обучающего сниппета 616 контента;
- третью входную группу 638, содержащую слова из третьего обучающего сниппета 618 контента;
- j -ю входную группу 640, содержащую слова из j -го обучающего сниппета 620 контента.

[128] В некоторых вариантах осуществления настоящей технологии каждое слово из обучающей текстовой входной последовательности 610 может быть связано с соответствующей позицией в обучающей текстовой входной последовательности 610. Дополнительно или в качестве альтернативы каждая входная группа из

45 последовательности входных групп может быть связана с соответствующей позицией в обучающей текстовой входной последовательности 610. В других вариантах осуществления изобретения предполагается, что данные, указывающие на обучающий запрос 642 и обучающую текстовую входную последовательность 610, могут вводиться

в алгоритм 400 MLA в векторном формате. Например, сервер 106 может формировать векторные представления для обучающего запроса 642 и обучающей текстовой входной последовательности 610 и использовать эти векторные представления в качестве входных данных для подсети 602 кодера.

5 [129] Подсеть 602 кодера способна формировать кодированное представление 650 обучающей текстовой входной последовательности 610. В некоторых вариантах осуществления настоящей технологии предполагается, что кодированное представление 650, сформированное подсетью 602 кодера, может состоять из одного или нескольких векторов, сформированных подсетью 602 кодера на итерации обучения и подлежащих
10 дальнейшей обработке подсетью 604 декодера.

[130] В общем случае подсеть 604 декодера способна формировать декодированное представление для обучающей текстовой входной последовательности 610. В приведенном примере это декодированное представление соответствует прогнозируемой текстовой выходной последовательности 652 сети 140 ANN. Следует отметить, что
15 прогнозируемая текстовая выходная последовательность 652 соответствует прогнозу, сформированному сетью 140 ANN, и представляет собой прогнозируемую сводку контента последовательности обучающих сниппетов контента из обучающей текстовой входной последовательности 610.

[131] Сервер 106 может сравнивать прогнозируемую текстовую выходную
20 последовательность 652 с заранее заданной текстовой выходной последовательностью 644 (прогноз сравнивается с целью). Предполагается, что заранее заданная текстовая выходная последовательность 644 может представлять собой оцененный человеком сниппет контента. Например, как описано выше, сервер 106 может получать запрос и соответствующий оцененный человеком сниппет контента из данных 330 оценщика-
25 человека, хранящихся в системе 120 базы данных, а затем использовать их в качестве обучающего запроса 642 и заранее заданной текстовой выходной последовательности 644, соответственно.

[132] Сервер 106 также способен формировать оценку 655 штрафа на основе сравнения прогнозируемой текстовой выходной последовательности 652 с заранее заданной
30 текстовой выходной последовательностью 644. На вид метрики штрафа, используемой сервером 106 для формирования оценки 655 штрафа на основе прогнозируемой текстовой выходной последовательности 652 и заранее заданной текстовой выходной последовательности 644, не накладывается каких-либо особых ограничений. Тем не менее, следует отметить, что оценка 655 штрафа указывает на различие между
35 прогнозируемой текстовой выходной последовательностью 652 и заранее заданной текстовой выходной последовательностью 644. Также можно сказать, что оценка 655 штрафа может указывать на степень сходства или различия прогнозируемой текстовой выходной последовательности 652 и заранее заданной текстовой выходной последовательности 644.

40 [133] На итерации обучения сервер 106 способен обучать сеть 140 ANN на основе оценки 655 штрафа. Например, сервер 106 может использовать метод обратного распространения для корректировки сети 140 ANN на основе оценки 655 штрафа. Тем не менее для корректировки сети 140 ANN с целью улучшения прогнозов также могут быть использованы другие способы обучения в сочетании с оценкой 655 штрафа.
45 Независимо от используемого конкретного способа обучения, можно сказать, что в результате выполнения итерации обучения сервером 106 сеть 140 ANN обучается формированию для запросов прогнозируемых текстовых выходных последовательностей, схожих с заранее заданными текстовыми выходными

последовательностями (т.е. формированию прогнозов, более близких к соответствующим целям). С учетом того, что заранее заданные текстовые выходные последовательности могут представлять собой оцененные людьми сниппеты контента из системы 120 базы данных, сеть 140 ANN таким образом может быть обучена формированию для запросов прогнозируемых текстовых выходных последовательностей, схожих с оцененными людьми сниппетами контента (например, с краткими сниппетами контента).

[134] Как описано выше, подсеть 602 кодера может применять конкретный механизм внимания в отношении обучающей текстовой входной последовательности 610 с целью формирования кодированного представления 650. В общем случае подсеть 602 кодера может применять маску 606 ограничения внимания в отношении обучающей текстовой входной последовательности 610 при формировании выходных данных вида «внимание», подлежащих использованию при формировании кодированного представления 650.

[135] В некоторых вариантах осуществления настоящей технологии можно сказать, что выходные данные вида «внимание» подсети 602 кодера могут указывать на контексты, в которых используются слова обучающей текстовой входной последовательности 610. Тем не менее, следует отметить, что при формировании выходных данных вида «внимание» для слова из входной группы обучающей текстовой входной последовательности 610 маска 606 ограничения внимания позволяет «принимать во внимание» только слова из этой входной группы так, чтобы выходные данные вида «внимание» формировались на основе контекста, полученного из этой входной группы, а не контекстов, полученных из других входных групп в обучающей текстовой входной последовательности 610. В результате подсеть 602 кодера может получать контент для слова из обучающей текстовой входной последовательности 610, зависящий только от слов из обучающего сниппета контента, к которому относится это слово, без учета слов из других обучающих сниппетов контента из обучающей текстовой входной последовательности 610.

[136] Например, можно предположить, что подсеть 602 кодера обрабатывает слово, связанное со второй входной группой 636, из второго обучающего сниппета 616 контента, содержащего три слова, как показано на представлении 690. Как показано, когда подсеть 602 кодера формирует для некоторого слова выходные данные 680 вида «внимание», маска 606 ограничения внимания применяется в отношении обучающей текстовой выходной последовательности 610 так, что подсеть 602 кодера может «принимать во внимание» только два других слова из той же входной группы (из второй входной группы 636) и «не принимает во внимание» слова из других входных групп обучающей текстовой входной последовательности 610. В результате можно сказать, что имеющая маску 606 ограничения внимания подсеть 602 кодера при формировании выходных данных вида «внимание» для слова из обучающей текстовой входной последовательности 610 использует контекст соответствующей входной группы (в данном случае - контекст, полученный из второй входной группы 636), а не контексты других входных групп.

[137] Также следует отметить, что в дополнение к использованию контекста соответствующей входной группы, подсеть 602 кодера может использовать контекст обучающего запроса 642. Предполагается, что можно сказать, что подсеть 602 кодера при формировании выходных данных вида «внимание» для слова из обучающей текстовой входной последовательности 610 использует контекст обучающего запроса 642 и контекст соответствующей входной группы (в данном случае - контекст, полученный из второй входной группы 636), а не контексты других входных групп. В по меньшей мере некоторых вариантах осуществления изобретения предполагается,

что контекст обучающего запроса 642 может представлять собой один или несколько терминов, содержащихся в обучающем запросе 642, без выхода за границы настоящей технологии.

[138] Вкратце, сеть 140 ANN может быть обучена во время основного этапа ее обучения на основе большого количества итераций, выполняемых подобно итерации обучения, представленной на фиг. 5. Как описано выше, сервер 106 может выполнять это большое количество итераций на основе соответствующих обучающих наборов данных.

[139] Следует отметить, что в по меньшей мере некоторых вариантах осуществления настоящей технологии предполагается, что во время основного этапа обучения сервер 106 может формировать и использовать для обучения сети 140 ANN модифицированные обучающие наборы данных.

[140] Для лучшей иллюстрации этого далее рассмотрен представленный на фиг. 5 обучающий набор данных, содержащий обучающую текстовую входную последовательность 610, обучающий запрос 642 и заранее заданную текстовую выходную последовательность 644. С учетом того, что обучающая текстовая входная последовательность 610 формируется в виде последовательности обучающих сниппетов контента, в некоторых не имеющих ограничительного характера вариантах осуществления настоящей технологии сервер 106 может формировать модифицированный обучающий набор данных путем (а) изменения порядка обучающих сниппетов контента в последовательности обучающих сниппетов контента обучающей текстовой входной последовательности 610 и (б) использования переупорядоченной таким образом обучающей текстовой входной последовательности (вместо последовательности 610), обучающего запроса 642 и заранее заданной текстовой выходной последовательности 644 в качестве модифицированного обучающего набора данных. Таким образом, предполагается, что сервер 106 может формировать модифицированный обучающий набор данных на основе обучающего набора данных путем формирования модифицированной обучающей текстовой входной последовательности, порядок входных групп в которой отличается от порядка входных групп в обучающей текстовой входной последовательности обучающего набора данных.

[141] Разработчики настоящей технологии установили, что благодаря обучению сети 140 ANN на основе обучающих наборов данных и на основе модифицированных наборов данных сеть 140 ANN может формировать прогнозы, меньше зависящие от порядка входных групп в текстовой входной последовательности, а значительно больше зависящие от контекстов соответствующих входных групп. В по меньшей мере некоторых вариантах осуществления настоящей технологии, если сеть 140 ANN используется для формирования сводки контента для страницы SERP на основе запроса и сниппетов контента результатов поиска на странице SERP, то обучение сети 140 ANN на основе обучающих наборов данных и на основе модифицированных наборов данных позволяет сети 140 ANN предоставлять сводки контента, меньше зависящие от порядка расположения результатов поиска на странице SERP. В результате сводки контента, сформированные сетью 140 ANN для набора результатов поиска, вероятно, будут схожи друг с другом (более согласованные сводки контента) вне зависимости от порядка ранжирования набора результатов поиска на странице SERP.

[142] Как описано выше, в некоторых не имеющих ограничительного характера вариантах осуществления настоящей технологии сервер 106 после основного этапа обучения сети 140 ANN может дополнительно выполнять этап ее обучения для предотвращения «галлюцинаций». Разработчики настоящей технологии установили,

что в некоторых не имеющих ограничительного характера вариантах осуществления изобретения сеть 140 ANN, обученная во время основного этапа ее обучения, может формировать «необычные» прогнозы для некоторых текстовых входных последовательностей этапа использования. Иными словами, сеть 140 ANN может «галлюцинировать» в том смысле, что прогноз, сформированный для некоторых текстовых входных последовательностей, может быть явно ошибочным.

[143] Для исправления этих ошибочных прогнозов в некоторых не имеющих ограничительного характера вариантах осуществления настоящей технологии сервер 106 может выполнять этап обучения для предотвращения «галлюцинаций», во время которого набор «наилучших» обучающих наборов данных выбирается для настройки качества прогнозирования сети 140 ANN. Этот набор наилучших обучающих наборов данных может быть заранее задан оценщиками-людьми. Сервер 106 также может заранее выбирать эти обучающие наборы данных как подмножество обучающих наборов данных, в которых заранее заданная текстовая выходная последовательность присутствует в одном или нескольких снippetах соответствующей обучающей текстовой входной последовательности.

[144] Обучив сеть 140 ANN, сервер 106 может применять сеть 140 ANN на этапе ее использования. Ниже со ссылкой на фиг. 6 описано возможное применение сервером 106 сети 140 ANN на этапе ее использования в по меньшей мере некоторых не имеющих ограничительного характера вариантах осуществления настоящей технологии.

Этап использования сети ANN

[145] На фиг. 6 приведено представление 700 возможного порядка применения сервером 106 сети 140 ANN на этапе ее использования. Как показано, звуковой сигнал 152 может быть зафиксирован и звуковое представление 702 речевого фрагмента пользователя 102 может быть сформировано электронным устройством 104. Звуковое представление 702 может быть обработано для формирования текстового представления 704 запроса пользователя. Следует отметить, что процедура 720 преобразования текста в речь (STT, Speech-To-Text) может быть выполнена электронным устройством 104 и/или сервером 106 различными способами без выхода за границы настоящей технологии.

[146] Сервер 106 может выполнять поиск в ответ на текстовое представление 704 запроса пользователя. Иными словами, сервер 106 может выполнять процедуру 730 поиска для получения одного или нескольких контент-ресурсов, релевантных запросу пользователя. В некоторых вариантах осуществления изобретения сервер 106 может использовать поисковую систему 130 для получения множества контент-ресурсов, релевантных запросу пользователя. Следует отметить, что сервер 106 также может использовать механизм 135 снippetов с целью формирования снippetов контента этапа использования для соответствующих контент-ресурсов из множества контент-ресурсов. Таким образом, сервер 106 может формировать страницу 706 SERP, содержащую множество результатов поиска, каждый из которых связан с соответствующим контент-ресурсом, релевантным запросу пользователя и имеющим соответствующий снippet контента этапа использования, сформированный на основе контента из соответствующего контент-ресурса.

[147] Сервер 106 может использовать информацию, сформированную поисковой системой 130 и механизмом 135 снippetов, для выполнения процедуры 740 формирования набора данных этапа использования. В общем случае при выполнении процедуры 740 формирования набора данных этапа использования сервер 106 может использовать информацию, сформированную поисковой системой 130 и механизмом 135 снippetов, для формирования текстовой входной последовательности 708 этапа

использования. Например, сервер 106 может формировать текстовую входную последовательность 708 этапа использования в виде последовательности сниппетов контента этапа использования, сформированных механизмом 135 сниппетов для соответствующих контент-ресурсов, релевантных запросу пользователя.

5 [148] Сервер 106 может вводить текстовое представление 704 запроса пользователя и текстовую входную последовательность 708 этапа использования в обученную сеть 140 ANN. Таким образом, можно сказать, что сервер 106 может выполнять процедуру 750 формирования сводки, во время которой итерация этапа использования сети 140 ANN выполняется в отношении введенных текстового
10 представления 704 запроса пользователя и текстовой входной последовательности 708 этапа использования. В результате обученная сеть 140 ANN выдает для сервера 106 текстовую выходную последовательность 710 этапа использования, представляющую собой (прогнозируемую) сводку контента последовательности сниппетов контента этапа использования (например, сформированных поисковой системой согласно
15 некоторым известным способам и/или сформированных механизмом 135 сниппетов, как описано выше) и подлежащую использованию в качестве ответа на запрос пользователя.

[149] Сервер 106 также может формировать звуковое представление 712 текстовой выходной последовательности 710. Звуковое представление 712 подлежит использованию
20 в качестве формируемого компьютером речевого фрагмента ответа. Предполагается, что процедура 750 преобразования текста в речь (TTS, Text-To-Speech) может выполняться сервером 106 и/или электронным устройством 104. Получив и/или сформировав звуковое представление 712, электронное устройство 104 может использовать это представление с целью формирования звукового сигнала 154 для
25 пользователя 102, например, с помощью одного или нескольких громкоговорителей.

[150] В некоторых не имеющих ограничительного характера вариантах осуществления настоящей технологии сервер 106 может выполнять способ 800, представленный на фиг. 7. Далее более подробно описаны различные шаги способа 800.

Шаг 802: ввод обучающего запроса и обучающей текстовой входной
30 последовательности.

[151] Способ 800 начинается с шага 802, на котором сервер 106 на итерации обучения сети 140 ANN может вводить обучающий запрос 642 и обучающую текстовую входную последовательность 610 в подсеть 602 кодера (см. фиг. 5). Обучающая текстовая входная последовательность 610 (а) формируется в виде последовательности обучающих
35 сниппетов контента (например, последовательности, содержащей сниппеты 612, 614, 616, 618, 620 контента) и (б) разделяется на последовательность входных групп (например, последовательность, содержащая входные группы 632, 634, 636, 638, 640). Следует отметить, что входная группа связана с обучающим сниппетом контента и содержит слова из обучающего сниппета контента.

40 [152] В некоторых вариантах осуществления изобретения обучающие сниппеты контента из обучающей текстовой входной последовательности могут быть сформированы сервером 106, использующим поисковую систему 130. Например, сервер 106 может выполнять поиск и использовать сниппеты контента, связанные с по меньшей мере некоторыми релевантными результатами поиска, в качестве соответствующих
45 обучающих сниппетов контента. В некоторых вариантах осуществления изобретения обучающие сниппеты контента могут формироваться известными в данной области техники способами.

[153] В некоторых вариантах осуществления изобретения сервер 106 может получать

обучающие сниппеты контента из соответствующих контент-ресурсов, релевантных обучающему запросу 642. Предполагается, что сервер 106 может получать обучающие сниппеты контента путем выполнения поиска с использованием поисковой системы 130 и обучающего запроса 642 в качестве входных данных. Также предполагается, что сервер 106 может получать обучающие сниппеты контента путем получения N наиболее релевантных сниппетов, связанных с N лучшими результатами поиска.

[154] В других вариантах осуществления изобретения сервер 106 может формировать обучающий сниппет контента из последовательности обучающих сниппетов контента. В некоторых вариантах осуществления изобретения сервер 106 может вводить обучающий запрос 642 и контент из обучающего контент-ресурса в алгоритм 400 MLA, способный выдавать прогнозируемый сниппет контента. Как описано выше, алгоритм 400 MLA может быть обучен сервером 106 формированию прогнозируемого сниппета контента, который подобен заранее заданной текстовой выходной последовательности и представляет собой обучающий сниппет контента, подлежащий использованию в качестве части обучающей текстовой входной последовательности 610.

Шаг 804: формирование кодированного представления обучающей текстовой входной последовательности.

[155] Способ 800 продолжается на шаге 804, на котором сервер 106 может с использованием подсети 602 кодера формировать кодированное представление 650 обучающей текстовой входной последовательности 610. Для формирования кодированного представления 650 сервер 106 может формировать выходные данные вида «внимание» (такие как выходные данные 680 вида «внимание») для соответствующих слов из обучающей текстовой входной последовательности 610 путем применения маски 606 ограничения внимания в отношении обучающей текстовой входной последовательности 610. Выходные данные вида «внимание» используются сервером 106 для формирования кодированного представления 650 обучающей текстовой входной последовательности 610.

[156] Следует отметить, что при формировании выходных данных 680 вида «внимание» для слова из входной группы 636 маска 606 ограничения внимания позволяет «принимать во внимание» только слова из входной группы 636 так, чтобы выходные данные 680 вида «внимание» формировались на основе контекста входной группы 636, а не контекстов других входных групп из обучающей текстовой входной последовательности 610.

[157] Также следует отметить, что в дополнение к использованию контекста соответствующей входной группы подсеть кодера может использовать контекст обучающего запроса 642. Предполагается, что можно сказать, что подсеть 602 кодера при формировании выходных данных вида «внимание» для слова из обучающей текстовой входной последовательности 610 использует контекст обучающего запроса 642 и контекст соответствующей входной группы (в данном случае - контекст, полученный из второй входной группы 636), а не контексты других входных групп. В по меньшей мере некоторых вариантах осуществления изобретения предполагается, что контекст обучающего запроса 642 может основываться на одном или нескольких терминах, содержащихся в обучающем запросе 642, без выхода за границы настоящей технологии.

Шаг 806: формирование декодированного представления обучающей текстовой входной последовательности.

[158] Способ 800 продолжается на шаге 806, на котором сервер 106 с использованием подсети 604 декодера может формировать для обучающей текстовой входной

последовательности 610 декодированное представление 652, соответствующее прогнозируемой текстовой выходной последовательности сети 140 ANN.

Предполагается, что декодированное представление 652 может выдаваться сетью 140 ANN в векторном формате без выхода за границы настоящей технологии.

5 [159] В некоторых вариантах осуществления изобретения предполагается, что подсеть 604 декодера может использовать собственный механизм внимания. Например, механизм внимания подсети 604 декодера может быть реализован подобно механизму внимания декодера из трансформерной сети. В некоторых случаях можно сказать, что механизм
10 внимания подсети 604 декодера может быть реализован в соответствии с известными в данной области техники вариантами для трансформерных сетей.

Шаг 808: формирование оценки штрафа для данной итерации обучения.

[160] Способ 800 продолжается на шаге 808, на котором сервер 106 может формировать оценку 655 штрафа для итерации обучения путем сравнения прогнозируемой текстовой выходной последовательности 652 с заранее заданной
15 текстовой выходной последовательностью 644. Заранее заданная текстовая выходная последовательность представляет собой заранее заданный ответ на обучающий запрос 642.

[161] Следует отметить, что в некоторых вариантах осуществления изобретения сервер 106 дополнительно может получать заранее заданную текстовую выходную
20 последовательность 644. В некоторых вариантах осуществления изобретения заранее заданная текстовая выходная последовательность 644 может быть получена от оценщика-человека (например, сформирована оценщиком-человеком и сохранена в системе 120 базы данных, как описано выше). В некоторых вариантах осуществления изобретения сервер 106 также может получать указание на контент-ресурс,
25 использованный оценщиком-человеком для формирования заранее заданной текстовой выходной последовательности 644. В некоторых вариантах осуществления изобретения сервер 106 может получать из системы 120 базы данных указание на контент-ресурс, который может быть связан с заранее заданной текстовой выходной последовательностью 644 из оцененного человеком набора данных. В одном не имеющем
30 ограничительного характера примере контент-ресурс может быть использован для проверки заранее заданной текстовой выходной последовательности 644. Также предполагается, что контент-ресурс может представлять собой сетевой ресурс, соответствующий обучающему запросу 642 и полученный путем выполнения поиска с использованием поисковой системы 130, без выхода за границы настоящей технологии.

35 Шаг 810: корректировка сети ANN на основе оценки штрафа.

[162] Способ 800 продолжается на шаге 810, на котором сервер 106 может корректировать сеть 140 ANN на основе оценки 655 штрафа. В одном не имеющем ограничительного характера примере сервер 106 для корректировки сети 140 ANN может применять метод обратного распространения в отношении сети 140 ANN на
40 основе оценки 655 штрафа.

[163] В некоторых вариантах осуществления настоящей технологии сервер 106 может выполнять большое количество итераций обучения подобно итерации обучения, представленной на фиг. 5. Например, на итерации обучения сеть 140 ANN обучается на основе обучающего набора данных, содержащего (а) обучающий запрос 642, (б)
45 обучающую текстовую входную последовательность 610 и (в) заранее заданную текстовую выходную последовательность 644. В некоторых вариантах осуществления настоящей технологии сервер 106 также может на другой итерации обучения обучать сеть 140 ANN на основе другого обучающего набора данных (модифицированного

обучающего набора данных). Например, другой обучающий набор данных может содержать (а) обучающий запрос 642, (б) модифицированную обучающую текстовую входную последовательность и (в) заранее заданную текстовую выходную последовательность 644, при этом входные группы в модифицированной обучающей текстовой входной последовательности расположены в порядке, отличном от порядка входных групп в обучающей текстовой входной последовательности 610.

[164] В некоторых вариантах осуществления настоящей технологии сервер 106 может выполнять представленную на фиг. 5 итерацию обучения во время первого (и/или основного) этапа обучения сети 140 ANN. Тем не менее сервер 106 также может выполнять второй этап обучения сети 140 ANN (например, этап обучения для предотвращения «галлюцинаций»), включающий в себя второе множество итераций обучения. На другой итерации обучения из второго множества итераций обучения сервер 106 может обучать сеть 140 ANN на основе другого обучающего набора данных из множества заранее заданных обучающих наборов данных, подлежащих использованию при выполнении второго множества итераций обучения. Например, другой обучающий набор данных может быть выбран оценщиком-человеком.

[165] В некоторых вариантах осуществления настоящей технологии заранее заданная текстовая выходная последовательность 644 может быть определена оценщиком-человеком. Например, заранее заданная текстовая выходная последовательность 644 может быть получена из данных 330 оценщика-человека, хранящихся в системе 120 базы данных. Таким образом, оценщик-человек может определять заранее заданную текстовую выходную последовательность 644 из контента по меньшей мере одного из обучающих контент-ресурсов, связанных с обучающими сниппетами контента из обучающей текстовой входной последовательности 610.

[166] В других вариантах осуществления настоящей технологии сервер 106 также может выполнять этап использования сети 140 ANN. Например, сервер 106 может формировать текстовое представление 704 (см. фиг. 6) запроса на основе звукового представления 702 речевого фрагмента пользователя, указывающего на запрос пользователя 102, связанного с электронным устройством 104. Сервер 106 также может определять результаты поиска, соответствующие текстовому представлению 704 запроса, при этом результат поиска содержит сниппет контента и указывает на контент-ресурс, релевантный запросу. Для этого сервер 106 может выполнять поиск в ответ на текстовое представление 704 запроса пользователя. Иными словами, сервер 106 может выполнять процедуру 730 поиска для получения одного или нескольких контент-ресурсов, релевантных запросу пользователя. В некоторых вариантах осуществления изобретения сервер 106 может использовать поисковую систему 130 для получения множества контент-ресурсов, релевантных запросу пользователя. Следует отметить, что сервер 106 также может использовать механизм 135 сниппетов с целью формирования сниппетов контента этапа использования для соответствующих контент-ресурсов из множества контент-ресурсов. Таким образом, сервер 106 может формировать страницу 706 SERP, содержащую множество результатов поиска, каждый из которых связан с контент-ресурсом, релевантным запросу пользователя и имеющим сниппет контента этапа использования, сформированный на основе контента из соответствующего контент-ресурса.

[167] Сервер 106 также может формировать текстовую входную последовательность 708 в виде последовательности сниппетов контента на основе сниппетов контента результатов поиска. Сервер 106 также может вводить текстовую входную последовательность 708 и текстовое представление 704 запроса в сеть 140 ANN,

способную выдавать текстовую выходную последовательность 710. Текстовая выходная последовательность 710 может представлять собой сводку контента последовательности сниппетов контента, которая может использоваться в качестве ответа на запрос. Сервер 106 также может формировать звуковое представление 712 текстовой выходной последовательности 710, представляющее собой формируемый компьютером речевой фрагмент ответа. Сервер 106 также может предоставлять электронному устройству 104 звуковое представление 712 текстовой выходной последовательности.

[168] Для специалиста в данной области могут быть очевидными возможные изменения и усовершенствования описанных выше вариантов осуществления настоящей технологии. Предшествующее описание приведено лишь в иллюстративных целях, а не для ограничения объема изобретения. Объем охраны настоящей технологии определяется исключительно объемом приложенной формулы изобретения.

(57) Формула изобретения

1. Способ обучения содержащей подсеть кодера и подсеть декодера нейронной сети внимания (ANN) формированию текстовой выходной последовательности на основе запроса и текстовой входной последовательности, формируемой в виде последовательности сниппетов контента, полученных из соответствующих контент-ресурсов, релевантных запросу, при этом текстовая выходная последовательность представляет собой сводку контента последовательности сниппетов контента и подлежит использованию в качестве ответа на запрос, а способ выполняется сервером, содержащим сеть ANN, и на итерации обучения сети ANN включает в себя:

- ввод сервером обучающего запроса и обучающей текстовой входной последовательности в подсеть кодера, при этом обучающая текстовая входная последовательность (а) формируется в виде последовательности обучающих сниппетов контента и (б) разделяется на последовательность входных групп, а входная группа связана с обучающим сниппетом контента и содержит слова из этого обучающего сниппета контента;

- формирование сервером, использующим подсеть кодера, кодированного представления обучающей текстовой входной последовательности, включая формирование выходных данных вида «внимание» для соответствующих слов из обучающей текстовой входной последовательности путем применения маски ограничения внимания в отношении обучающей текстовой входной последовательности, а выходные данные вида «внимание» используются для формирования кодированного представления обучающей текстовой входной последовательности, при формировании выходных данных вида «внимание» для слова из входной группы маска ограничения внимания позволяет «принимать во внимание» только слова из этой входной группы так, чтобы выходные данные вида «внимание» формировались на основе контекста этой входной группы, а не контекстов других входных групп в обучающей текстовой входной последовательности;

- формирование сервером, использующим подсеть декодера, декодированного представления для обучающей текстовой входной последовательности, соответствующего прогнозируемой текстовой выходной последовательности;

- формирование сервером оценки штрафа для итерации обучения путем сравнения прогнозируемой текстовой выходной последовательности с заранее заданной текстовой выходной последовательностью, представляющей собой заранее заданный ответ на обучающий запрос; и

- корректировку сервером сети ANN на основе оценки штрафа.

2. Способ по п. 1, отличающийся тем, что формируемые выходные данные вида «внимание» дополнительно основаны на контексте из обучающего запроса.

3. Способ по п. 1, отличающийся тем, что на итерации обучения сеть ANN обучается на основе обучающего набора данных, содержащего (а) обучающий запрос, (б) обучающую текстовую входную последовательность и (в) заранее заданную текстовую выходную последовательность, при этом способ на другой итерации обучения дополнительно включает в себя обучение сети ANN на основе другого обучающего набора данных, содержащего (а) обучающий запрос, (б) модифицированную обучающую текстовую входную последовательность и (в) заранее заданную текстовую выходную последовательность, а входные группы в модифицированной обучающей текстовой входной последовательности расположены в порядке, отличном от порядка входных групп в обучающей текстовой входной последовательности.

4. Способ по п. 1, отличающийся тем, что он дополнительно включает в себя формирование обучающего сниппета контента из последовательности обучающих сниппетов контента, включая ввод сервером обучающего запроса и контента из обучающего контент-ресурса в алгоритм машинного обучения (MLA), способный выдавать прогнозируемый сниппет контента, при этом алгоритм MLA обучен формированию прогнозируемого сниппета контента так, чтобы прогнозируемый сниппет контента был подобен заранее заданной текстовой выходной последовательности и представлял собой обучающий сниппет контента.

5. Способ по п. 1, отличающийся тем, что на итерации обучения сеть ANN обучается на основе обучающего набора данных, содержащего (а) обучающий запрос, (б) обучающую текстовую входную последовательность и (в) заранее заданную текстовую выходную последовательность, при этом итерация обучения выполняется на первом этапе обучения сети ANN, включающем в себя первое множество итераций обучения, способ дополнительно включает в себя выполнение второго этапа обучения сети ANN, включающего в себя второе множество итераций обучения, и на другой итерации обучения из второго множества итераций обучения сеть ANN обучается на основе другого обучающего набора данных, который представляет собой набор из множества заранее заданных обучающих наборов данных, подлежащих использованию при выполнении второго множества итераций обучения, и выбирается оценщиком-человеком.

6. Способ по п. 1, отличающийся тем, что на итерации обучения сеть ANN обучается на основе обучающего набора данных, содержащего (а) обучающий запрос, (б) обучающую текстовую входную последовательность и (в) заранее заданную текстовую выходную последовательность, определенную оценщиком-человеком, при этом оценщик-человек определяет заранее заданную текстовую выходную последовательность из контента по меньшей мере одного из обучающих контент-ресурсов, связанных с обучающими сниппетами контента.

7. Способ по п. 1, отличающийся тем, что он дополнительно включает в себя:

- формирование сервером текстового представления запроса на основе звукового представления речевого фрагмента пользователя, указывающего на запрос пользователя, связанного с электронным устройством;
- определение сервером результатов поиска, соответствующих текстовому представлению запроса, при этом результат поиска содержит сниппет контента и указывает на контент-ресурс, релевантный запросу;
- формирование сервером текстовой входной последовательности в виде последовательности сниппетов контента на основе сниппетов контента результатов поиска;

- ввод сервером текстовой входной последовательности и текстового представления запроса в сеть ANN, способную выдавать текстовую выходную последовательность, представляющую собой сводку контента последовательности сниппетов контента и подлежащую использованию в качестве ответа на запрос;

5 - формирование сервером звукового представления текстовой выходной последовательности, представляющего собой формируемый компьютером речевой фрагмент ответа; и

- предоставление сервером электронному устройству, связанному с пользователем, звукового представления текстовой выходной последовательности.

10 8. Способ по п. 1, отличающийся тем, что корректировка включает в себя применение сервером метода обратного распространения в отношении сети ANN.

9. Способ по п. 1, отличающийся тем, что он дополнительно включает в себя получение сниппетов контента из соответствующих контент-ресурсов, релевантных запросу.

15 10. Способ по п. 9, отличающийся тем, что получение включает в себя выполнение поиска с использованием поисковой системы и запроса в качестве входных данных.

11. Способ по п. 10, отличающийся тем, что получение включает в себя получение N наиболее релевантных сниппетов, связанных с N лучшими результатами поиска.

12. Способ по п. 1, отличающийся тем, что он дополнительно включает в себя
20 получение заранее заданной текстовой выходной последовательности.

13. Способ по п. 12, отличающийся тем, что получение осуществляется от оценщика-человека.

14. Способ по п. 13, отличающийся тем, что он дополнительно включает в себя получение указания на контент-ресурс, использованный оценщиком-человеком для
25 формирования заранее заданной текстовой выходной последовательности.

15. Способ по п. 14, отличающийся тем, что контент-ресурс используется для проверки заранее заданной текстовой выходной последовательности.

16. Способ по п. 15, отличающийся тем, что контент-ресурс представляет собой сетевой ресурс, соответствующий обучающему запросу и полученный путем выполнения
30 поиска с использованием поисковой системы.

17. Сервер для обучения содержащей подсеть кодера и подсеть декодера нейронной сети внимания (ANN) формированию текстовой выходной последовательности на основе запроса и текстовой входной последовательности, сформированной в виде последовательности сниппетов контента, полученных из соответствующих контент-ресурсов, релевантных запросу, при этом текстовая выходная последовательность
35 представляет собой сводку контента последовательности сниппетов контента и подлежит использованию в качестве ответа на запрос, а сервер содержит сеть ANN и выполнен с возможностью на итерации обучения сети ANN:

- ввода обучающего запроса и обучающей текстовой входной последовательности
40 в подсеть кодера, при этом обучающая текстовая входная последовательность (а) сформирована в виде последовательности обучающих сниппетов контента и (б) разделена на последовательность входных групп, а входная группа связана с обучающим сниппетом контента и содержит слова из этого обучающего сниппета контента;

- формирования с использованием подсети кодера кодированного представления
45 обучающей текстовой входной последовательности, включая формирование выходных данных вида «внимание» для соответствующих слов из обучающей текстовой входной последовательности путем применения маски ограничения внимания в отношении обучающей текстовой входной последовательности, при этом выходные данные вида

«внимание» используются для формирования кодированного представления обучающей текстовой входной последовательности, а при формировании выходных данных вида «внимание» для слова из входной группы маска ограничения внимания позволяет «принимать во внимание» только слова из этой входной группы так, чтобы выходные данные вида «внимание» формировались на основе контекста этой входной группы, а не контекстов других входных групп в обучающей текстовой входной последовательности;

- формирования с использованием подсети декодера декодированного представления для обучающей текстовой входной последовательности, соответствующего прогнозируемой текстовой выходной последовательности;

- формирования оценки штрафа для итерации обучения путем сравнения прогнозируемой текстовой выходной последовательности с заранее заданной текстовой выходной последовательностью, представляющей собой заранее заданный ответ на обучающий запрос; и

- корректировки сети ANN на основе оценки штрафа.

18. Сервер по п. 17, отличающийся тем, что формируемые выходные данные вида «внимание» дополнительно основаны на контексте из обучающего запроса.

19. Сервер по п. 17, отличающийся тем, что на итерации обучения сеть ANN обучается на основе обучающего набора данных, содержащего (а) обучающий запрос, (б) обучающую текстовую входную последовательность и (в) заранее заданную текстовую выходную последовательность, и сервер дополнительно способен на другой итерации обучения обучать сеть ANN на основе другого обучающего набора данных, содержащего (а) обучающий запрос, (б) модифицированную обучающую текстовую входную последовательность и (в) заранее заданную текстовую выходную последовательность, при этом входные группы в модифицированной обучающей текстовой входной последовательности расположены в порядке, отличном от порядка входных групп в обучающей текстовой входной последовательности.

20. Сервер по п. 17, отличающийся тем, что он дополнительно выполнен с возможностью формирования обучающего снippets контента из последовательности обучающих снippets контента, включая ввод обучающего запроса и контента из обучающего контент-ресурса в алгоритм MLA, способный выдавать прогнозируемый снippet контента, при этом алгоритм MLA обучен формированию прогнозируемого снippets контента так, чтобы прогнозируемый снippet контента был подобен заранее заданной текстовой выходной последовательности и представлял собой обучающий снippet контента.

21. Сервер по п. 17, отличающийся тем, что на итерации обучения сеть ANN обучается на основе обучающего набора данных, содержащего (а) обучающий запрос, (б) обучающую текстовую входную последовательность и (в) заранее заданную текстовую выходную последовательность, при этом итерация обучения выполняется на первом этапе обучения сети ANN, включающем в себя первое множество итераций обучения, сервер дополнительно выполнен с возможностью выполнения второго этапа обучения сети ANN, включающего в себя второе множество итераций обучения, и на другой итерации обучения из второго множества итераций обучения сеть ANN обучается на основе другого обучающего набора данных, который представляет собой набор из множества заранее заданных обучающих наборов данных, подлежащих использованию при выполнении второго множества итераций обучения, и выбирается оценщиком-человеком.

22. Сервер по п. 17, отличающийся тем, что на итерации обучения сеть ANN обучается

на основе обучающего набора данных, содержащего (а) обучающий запрос, (б) обучающую текстовую входную последовательность и (в) заранее заданную текстовую выходную последовательность, определенную оценщиком-человеком, при этом оценщик-человек определяет заранее заданную текстовую выходную последовательность из контента по меньшей мере одного из обучающих контент-ресурсов, связанных с обучающими сниппетами контента.

23. Сервер по п. 17, отличающийся тем, что он дополнительно выполнен с возможностью:

- формирования текстового представления запроса на основе звукового представления речевого фрагмента пользователя, указывающего на запрос пользователя, связанного с электронным устройством;
- определения результатов поиска, соответствующих текстовому представлению запроса, при этом результат поиска содержит сниппет контента и указывает на контент-ресурс, релевантный запросу;
- формирования текстовой входной последовательности в виде последовательности сниппетов контента на основе сниппетов контента результатов поиска;
- ввода текстовой входной последовательности и текстового представления запроса в сеть ANN, способную выдавать текстовую выходную последовательность, представляющую собой сводку контента последовательности сниппетов контента и подлежащую использованию в качестве ответа на запрос;
- формирования звукового представления текстовой выходной последовательности, представляющего собой формируемый компьютером речевой фрагмент ответа; и
- предоставления электронному устройству, связанному с пользователем, звукового представления текстовой выходной последовательности.

24. Сервер по п. 17, отличающийся тем, что корректировка включает в себя применение сервером метода обратного распространения в отношении сети ANN.

25. Сервер по п. 17, отличающийся тем, что он дополнительно выполнен с возможностью получения сниппетов контента из соответствующих контент-ресурсов, релевантных запросу.

26. Сервер по п. 25, отличающийся тем, что получение включает в себя выполнение поиска с использованием поисковой системы и запроса в качестве входных данных.

27. Сервер по п. 26, отличающийся тем, что получение включает в себя получение N наиболее релевантных сниппетов, связанных с N лучшими результатами поиска.

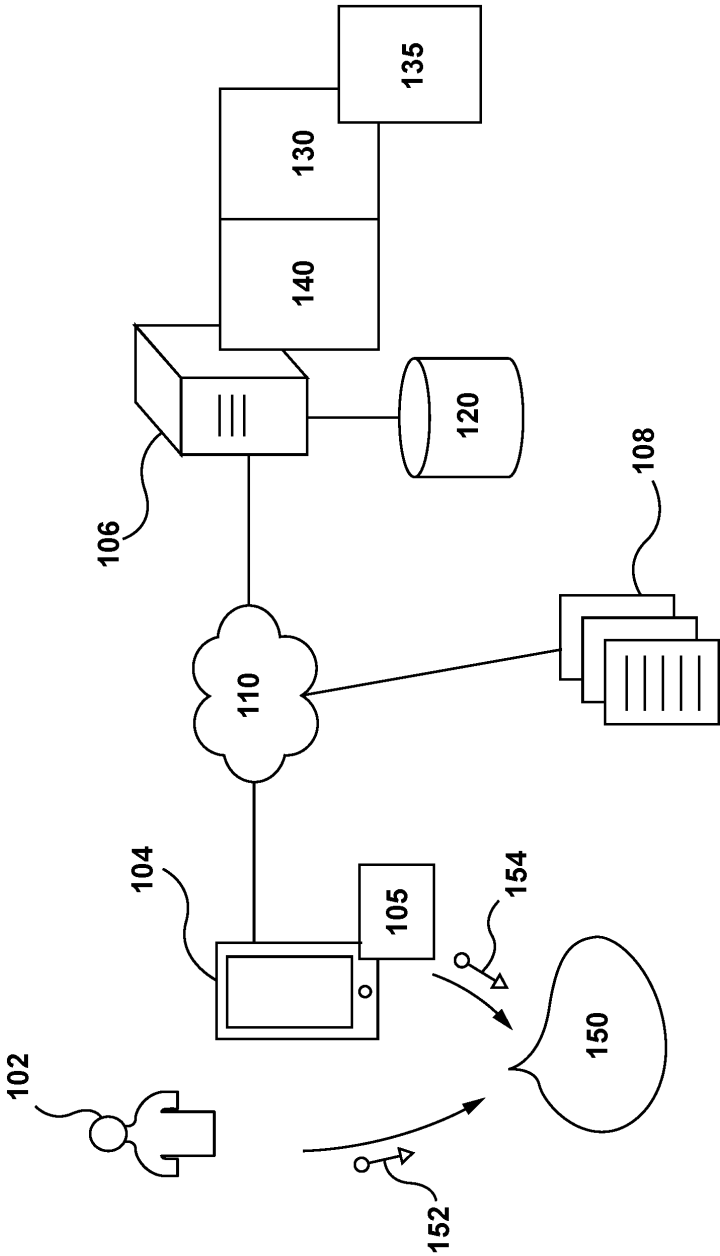
28. Сервер по п. 17, отличающийся тем, что он дополнительно выполнен с возможностью получения заранее заданной текстовой выходной последовательности.

29. Сервер по п. 28, отличающийся тем, что он выполнен с возможностью получения заранее заданной текстовой выходной последовательности от оценщика-человека.

30. Сервер по п. 29, отличающийся тем, что он дополнительно выполнен с возможностью получения указания на контент-ресурс, использованный оценщиком-человеком для формирования заранее заданной текстовой выходной последовательности.

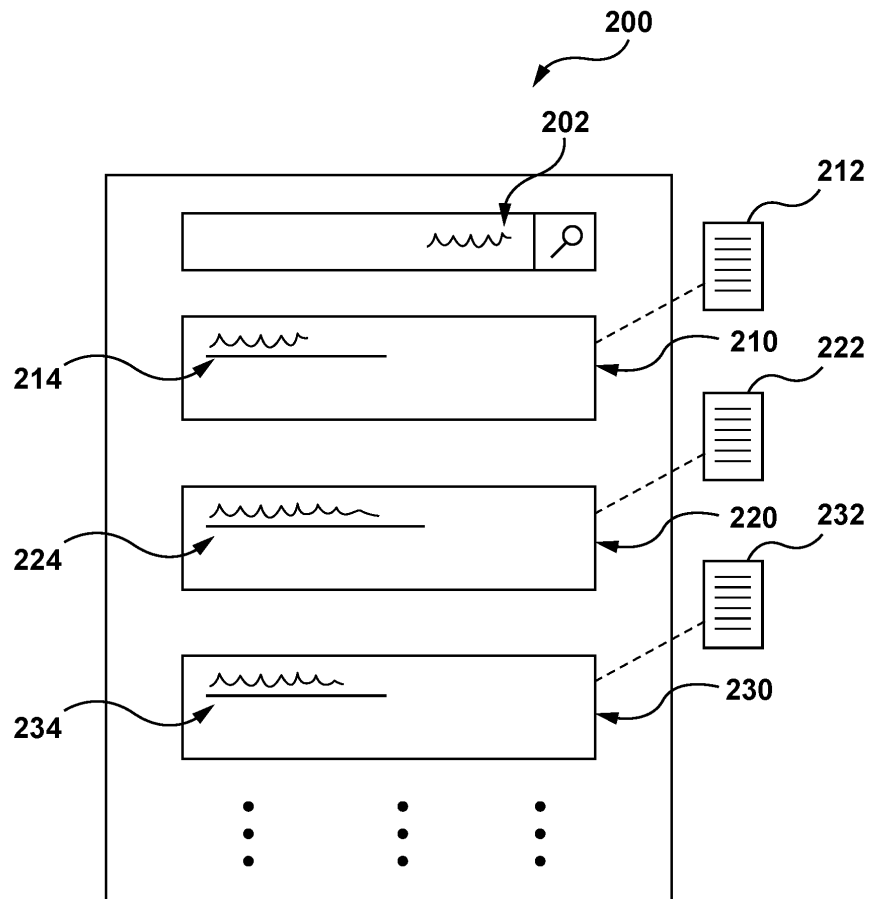
31. Сервер по п. 30, отличающийся тем, что контент-ресурс используется для проверки заранее заданной текстовой выходной последовательности.

32. Сервер по п. 31, отличающийся тем, что контент-ресурс представляет собой сетевой ресурс, соответствующий обучающему запросу и полученный путем выполнения поиска с использованием поисковой системы.



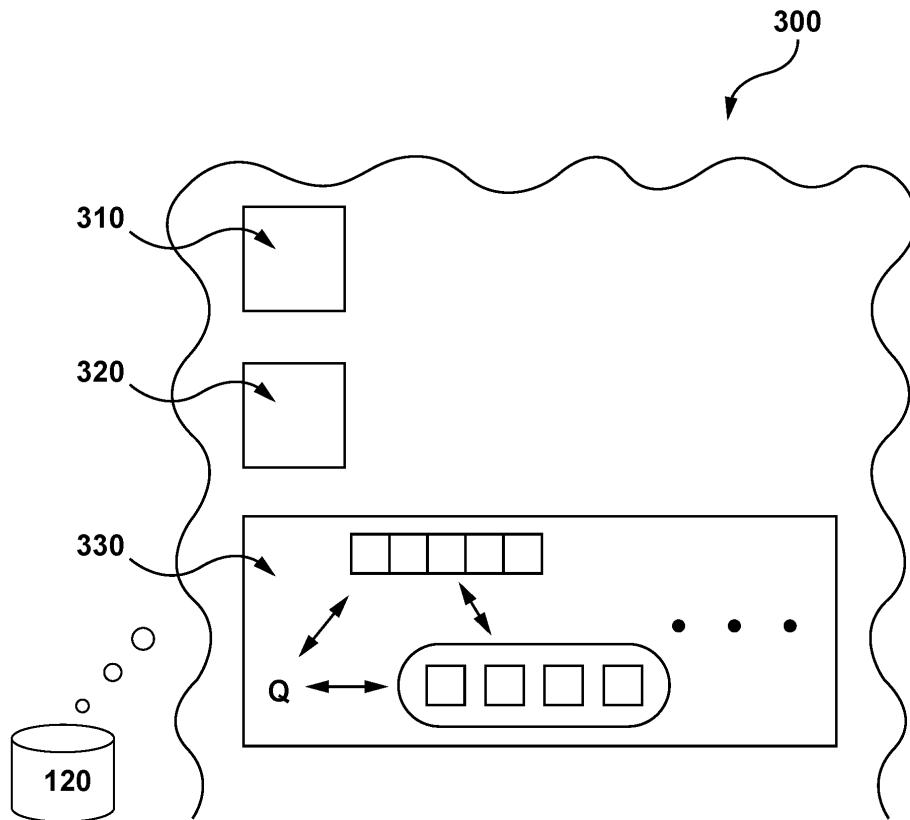
Фиг. 1

2/7



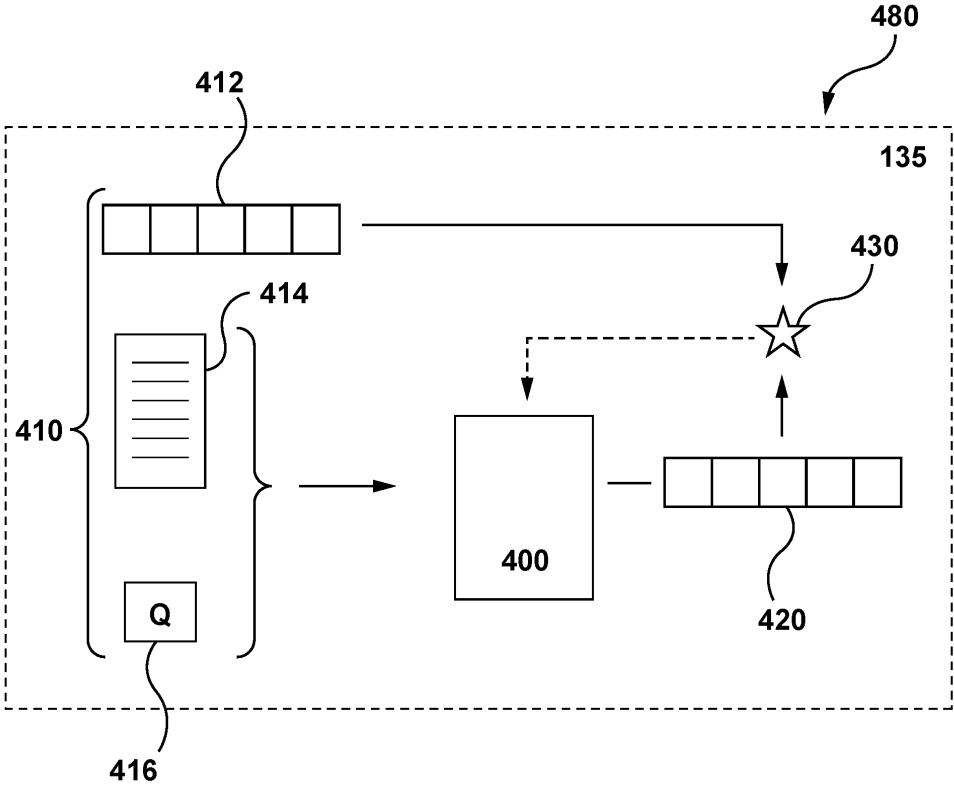
Фиг. 2

3/7



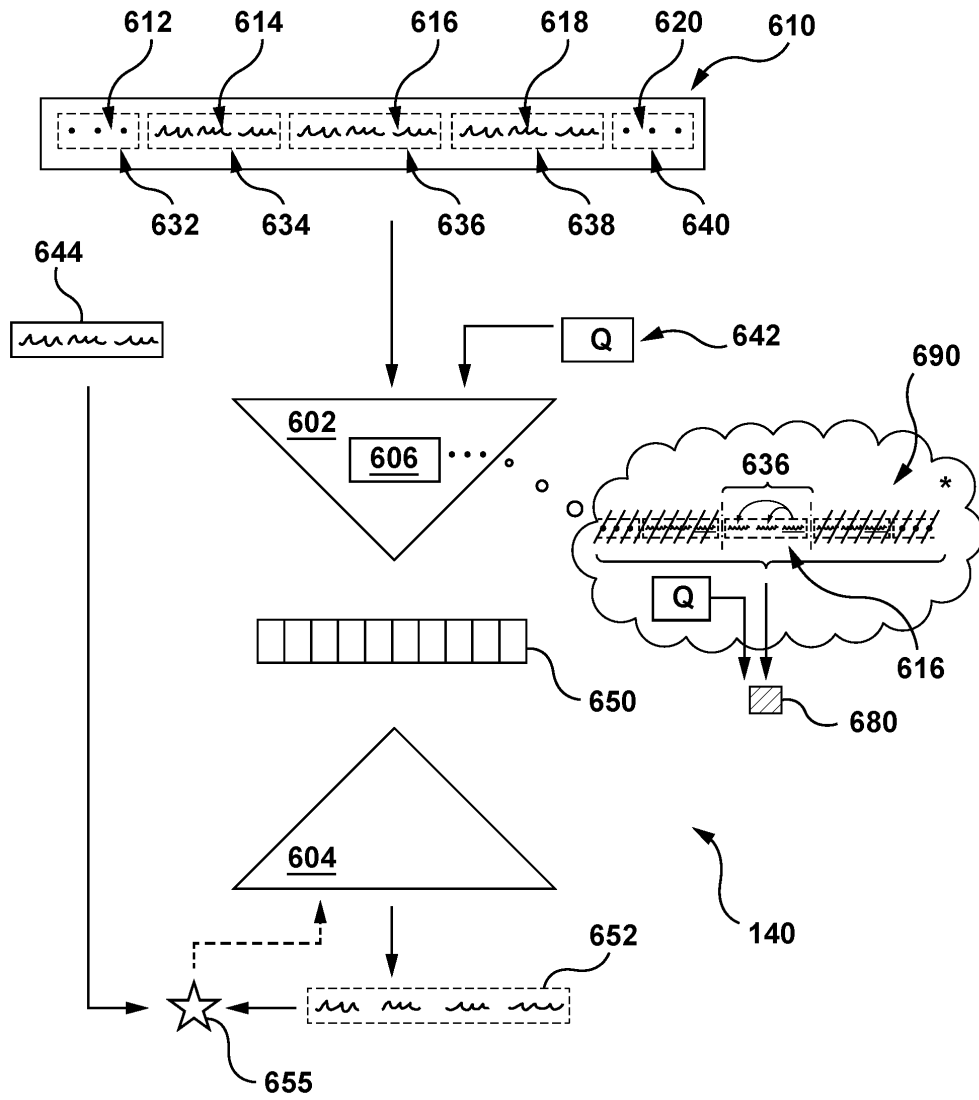
Фиг. 3

4/7



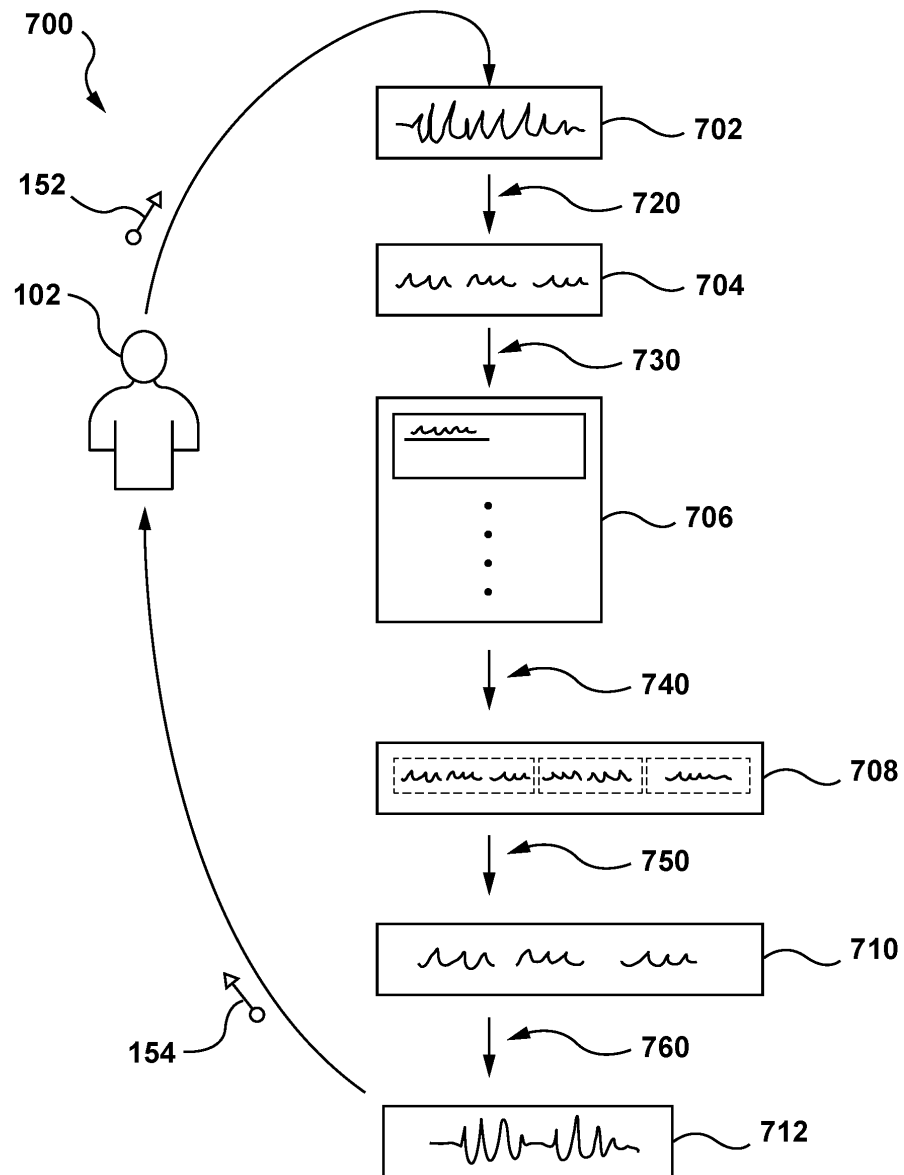
Фиг. 4

5/7



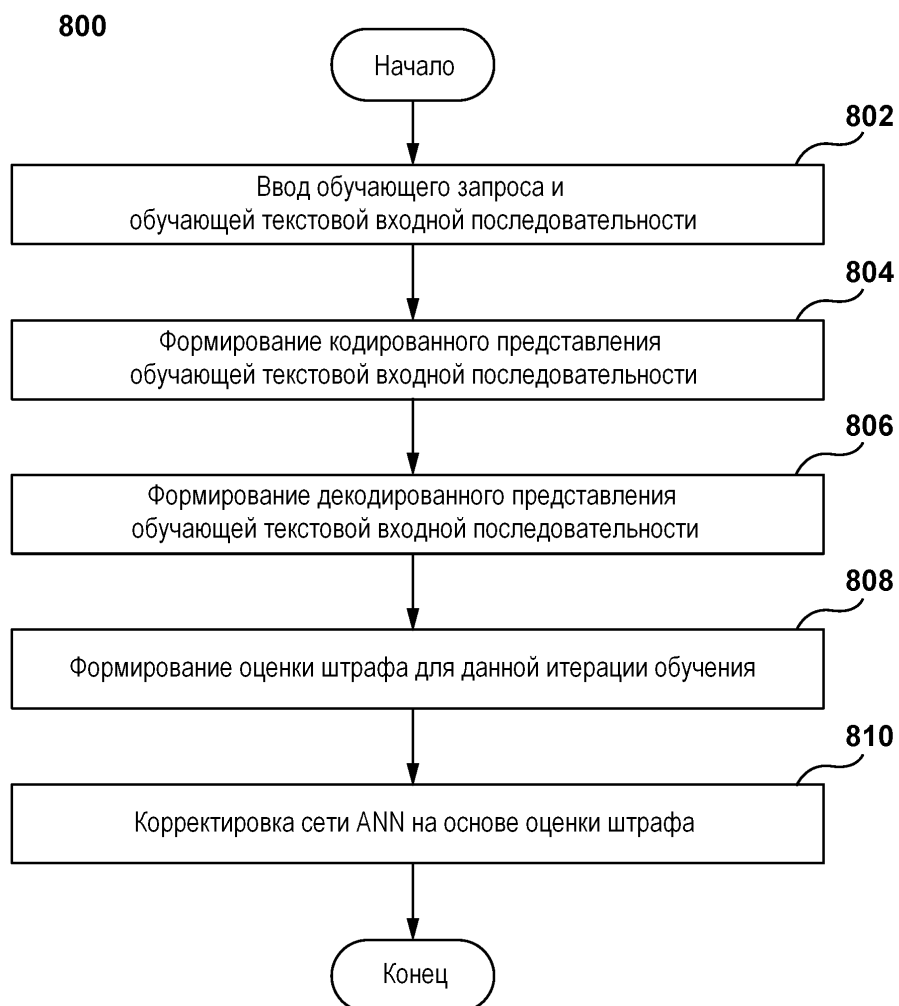
Фиг. 5

6/7



Фиг. 6

7/7



Фиг. 7