РОССИЙСКАЯ ФЕДЕРАЦИЯ

(19) **RU** (11) **2 691 855**(13) **C1**

(51) МПК
G06K 9/66 (2006.01)
G06F 17/27 (2006.01)
G06F 17/28 (2006.01)

ФЕДЕРАЛЬНАЯ СЛУЖБА
ПО ИНТЕЛЛЕКТУАЛЬНОЙ СОБСТВЕННОСТИ

(12) ОПИСАНИЕ ИЗОБРЕТЕНИЯ К ПАТЕНТУ

(52) СПК
G06F 17/271 *(2018.08);* ***G06F 17/2755*** *(2018.08);* ***G06F 17/2765*** *(2018.08);* ***G06F 17/2785*** *(2018.08);* ***G06F 17/28*** *(2018.08);* ***G06K 9/66*** *(2018.08)*

(21)(22) Заявка: **2018110387, 23.03.2018**

(24) Дата начала отсчета срока действия патента:
23.03.2018

Дата регистрации:
18.06.2019

Приоритет(ы):
(22) Дата подачи заявки: **23.03.2018**

(45) Опубликовано: **18.06.2019** Бюл. № **17**

Адрес для переписки:
127273, Москва, а/я 20, ООО "Аби Продакшн", для Марея С.В.

(72) Автор(ы):
Мацкевич Степан Евгеньевич (RU),
Булгаков Илья Александрович (RU)

(73) Патентообладатель(и):
Общество с ограниченной ответственностью "Аби Продакшн" (RU)

(56) Список документов, цитированных в отчете о поиске: **RU 2640718 C1, 11.01.2018. RU 2628431 C1, 16.08.2017. RU 2640297 C2, 27.12.2017. US 2017/0293607 A1, 12.10.2017. US 2016/0162456 A1, 09.06.2016.**

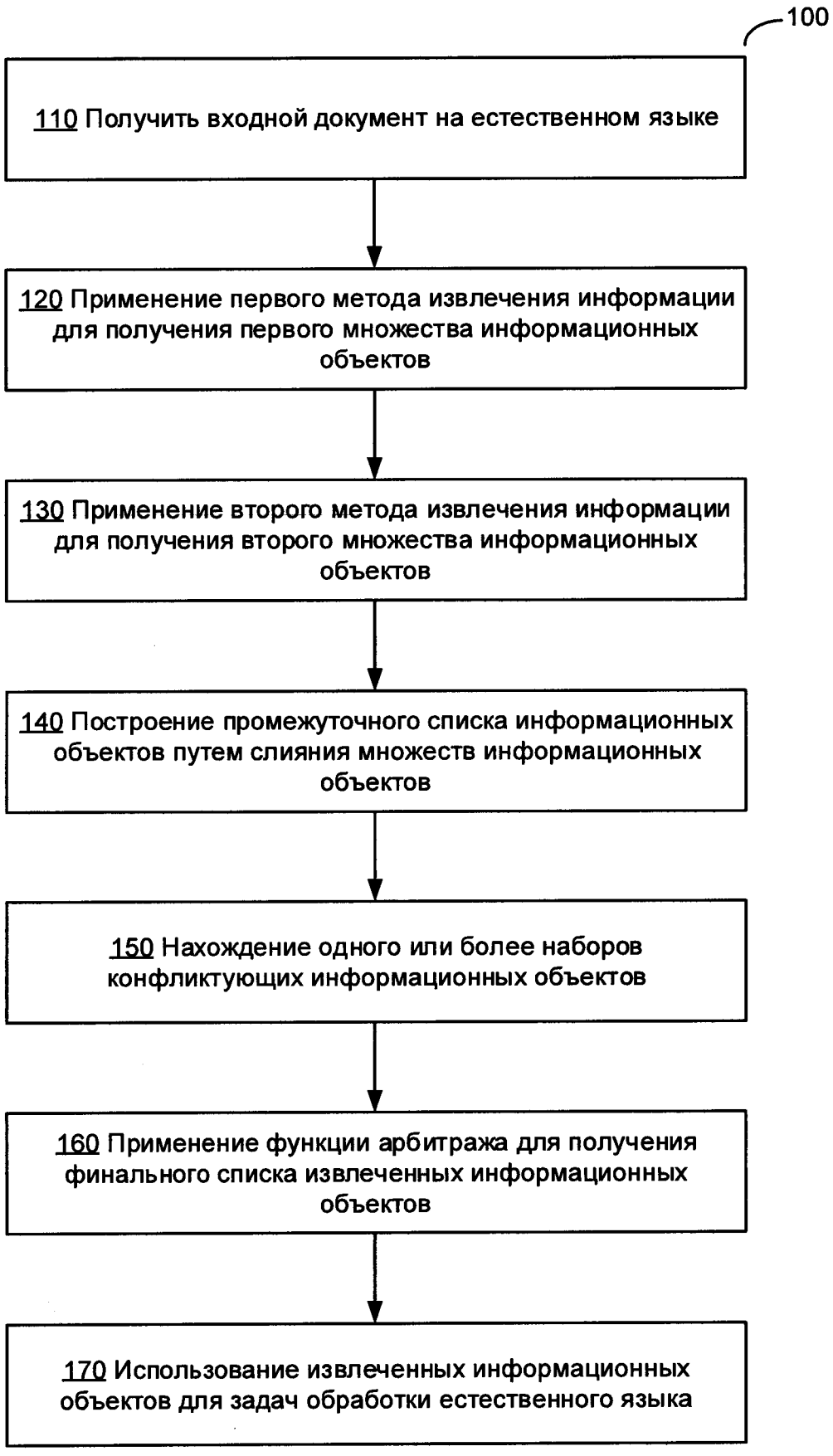
(54) ОБУЧЕНИЕ КЛАССИФИКАТОРОВ, ИСПОЛЬЗУЕМЫХ ДЛЯ ИЗВЛЕЧЕНИЯ ИНФОРМАЦИИ ИЗ ТЕКСТОВ НА ЕСТЕСТВЕННОМ ЯЗЫКЕ

(57) Реферат:

Изобретение относится к системе и способам извлечения информации из текстов на естественном языке. Техническим результатом является повышение эффективности и качества извлечения информации из текстов на естественном языке. Способ извлечения информации из текстов на естественном языке включает: обучение классификатора извлечения информации для извлечения первого множества информационных объектов из текста на естественном языке, причем определение классификатора извлечения информации включает один или более гиперпараметров; получение списка извлеченных информационных объектов путем выполнения функции арбитража конфликтов по отношению к множеству конфликтующих информационных объектов; изменение значений гиперпараметров классификатора извлечения информации; и оптимизацию показателя качества извлечения информации для списка извлеченных информационных объектов путем итеративного повторения операций обучения классификатора извлечения информации, выполнения функции арбитража конфликтов и изменения значений гиперпараметров. 4 н. и 21 з.п. ф-лы, 16 ил.

RU 2691855 C1

RU 2691855 C1



Фиг. 1

RU 2691855 C1

RUSSIAN FEDERATION

FEDERAL SERVICE
FOR INTELLECTUAL PROPERTY

(19) **RU** (11) **2 691 855**(13) **C1**

(51) Int. Cl.
G06K 9/66 (2006.01)
G06F 17/27 (2006.01)
G06F 17/28 (2006.01)

(12) ABSTRACT OF INVENTION

(52) CPC
G06F 17/271 *(2018.08);* ***G06F 17/2755*** *(2018.08);* ***G06F 17/2765*** *(2018.08);* ***G06F 17/2785*** *(2018.08);* ***G06F 17/28*** *(2018.08);* ***G06K 9/66*** *(2018.08)*

(21)(22) Application: **2018110387, 23.03.2018**

(24) Effective date for property rights:
23.03.2018

Registration date:
18.06.2019

Priority:
(22) Date of filing: **23.03.2018**

(45) Date of publication: **18.06.2019** Bull. № **17**

Mail address:
127273, Moskva, a/ya 20, OOO "Abi Prodakshn", dlya Mareya S.V.

(72) Inventor(s):
Matskevich Stepan Evgenevich (RU),
Bulgakov Ilya Aleksandrovich (RU)

(73) Proprietor(s):
Obshchestvo s ogranichennoj otvetstvennostyu "Abi Prodakshn" (RU)

RU 2691855 C1

(54) TRAINING CLASSIFIERS USED TO EXTRACT INFORMATION FROM NATURAL LANGUAGE TEXTS

(57) Abstract:

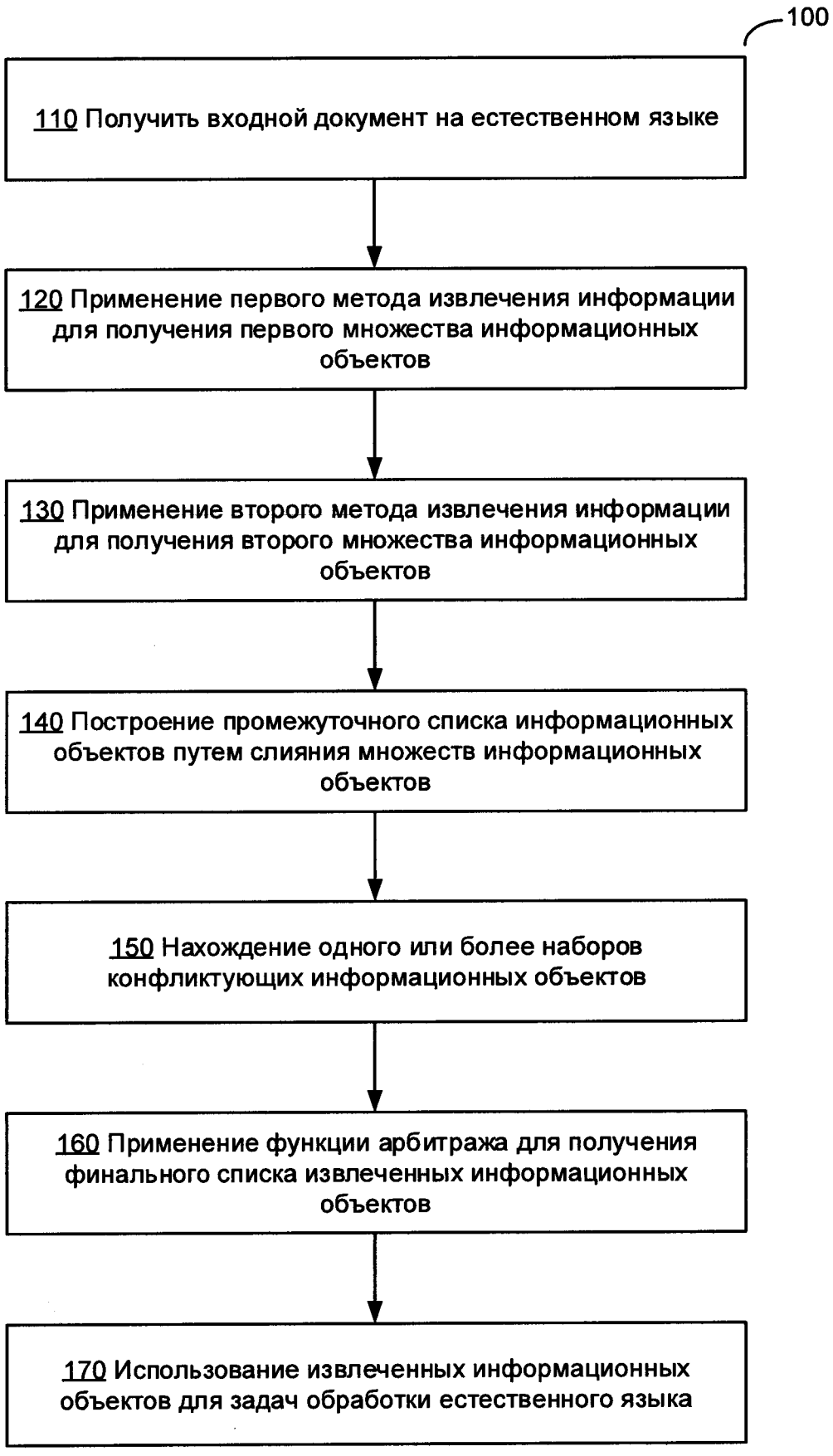
FIELD: data processing.

SUBSTANCE: invention relates to a system and methods of extracting information from natural language texts. Method of extracting information from natural language texts includes: training information extraction classifier to extract first plurality of information objects from text in natural language, wherein determination of information extraction classifier includes one or more hyperparameters; obtaining a list of extracted information objects by performing a conflict arbitration function with respect to a plurality of conflicting information objects; changing values of hyperparameters of information extraction classifier; and optimizing the information extraction quality factor for the list of extracted information objects by iterative repetition of the information extraction classifier training operations, performing the conflict arbitration function and changing the hyperparameter values.

EFFECT: high efficiency and quality of extracting information from natural language texts.

25 cl, 16 dwg

RU 2691855 C1



Фиг. 1

RU 2691855 C1

RU 2691855 C1

ОБЛАСТЬ ТЕХНИКИ

[0001] Настоящее изобретение в целом относится к вычислительным системам, а точнее - к системам и способам обработки естественного языка.

УРОВЕНЬ ТЕХНИКИ

[0002] Извлечение информации может включать анализ текста на естественном языке для выявления и классификации информационных объектов в соответствии с заданным множеством категорий (например, имена лиц, организаций, мест, выражения времени, количества, денежных сумм, процентов и т.д.). Извлечение информации может дополнительно включать определение отношений между распознанными именованными сущностями и (или) информационными объектами.

РАСКРЫТИЕ ИЗОБРЕТЕНИЯ

[0003] В соответствии с одним или более вариантами реализации настоящего изобретения пример способа обучения классификатора, используемого для извлечения информации из текстов на естественном языке может включать: обучение классификатора извлечения информации для извлечения первого множества информационных объектов из текста на естественном языке, причем определение классификатора извлечения информации включает один или более гиперпараметров; получение итогового списка извлеченных информационных объектов путем выполнения функции «арбитража» конфликтов по отношению к множеству конфликтующих информационных объектов, так, что первый информационный объект из набора конфликтующих информационных объектов первому множеству информационных объектов, и второй информационный объект из набора конфликтующих информационных объектов принадлежит второму множеству информационных объектов, извлеченных из текста на естественном языке; изменение значений гиперпараметров классификатора извлечения информации; и оптимизация показателя качества извлечения информации для итогового списка извлеченных информационных объектов путем итеративного повторения операций обучения классификатора извлечения информации, выполнения функции «арбитража» конфликтов и изменения значений гиперпараметров.

[0004] В соответствии с одним или более вариантами реализации настоящего изобретения другой пример способа обучения классификатора, используемого для извлечения информации из текстов на естественном языке может включать: извлечение, посредством первого метода извлечения информации, первого множества информационных объектов из текста на естественном языке; извлечение, посредством второго метода извлечения информации, второго множества информационных объектов из текста на естественном языке; изменение значений одного или нескольких гиперпараметров классификатора арбитража конфликтов; обучение классификатора арбитража конфликтов, который создает список извлеченных информационных объектов посредством выполнения функции арбитража конфликтов в отношении набора конфликтующих информационных объектов, так что первый информационный объект множества конфликтующих информационных объектов принадлежит первому множеству информационных объектов второй информационный объект множества конфликтующих информационных объектов принадлежит второму множеству информационных объектов; оптимизацию метрики качества извлечения информации, отражающую качество списка извлеченных информационных объектов, итеративно повторяя операции извлечения первого множества информационных объектов, извлечение второго множества информационных объектов, изменение значений гиперпараметров и обучение классификатора арбитража конфликтов.

[0005] В соответствии с одним или более вариантами реализации настоящего изобретения пример системы извлечения информации из текста на естественном языке может включать: запоминающее устройство и процессор, соединенный с запоминающим устройством. Этот процессор может быть выполнен с возможностью: обучения классификаторов извлечения информации для извлечения первого множества информационных объектов из текста на естественном языке, причем определение классификатора извлечения информации включает один или более гиперпараметров; получения итогового списка извлеченных информационных объектов путем выполнения функции «арбитража» конфликтов по отношению к множеству конфликтующих информационных объектов, например, первого информационного объекта из набора конфликтующих информационных объектов, принадлежащего к первому множеству информационных объектов, и второго информационного объекта из набора конфликтующих информационных объектов, принадлежащего ко второму множеству информационных объектов, извлеченных из текста на естественном языке; изменения значений гиперпараметров классификатора извлечения информации; и оптимизации показателя качества извлечения информации для итогового списка извлеченных информационных объектов путем итеративного повторения операций обучения классификатора извлечения информации, выполнения функции «арбитража» конфликтов и изменения значений гиперпараметров.

[0006] В соответствии с одним или более вариантами реализации настоящего изобретения пример машиночитаемого постоянного носителя данных может содержать исполняемые команды, которые при выполнении вычислительной системой вызывают следующие действия вычислительной системы: обучение классификаторов извлечения информации для извлечения первого множества информационных объектов из текста на естественном языке, причем определение классификатора извлечения информации включает один или более гиперпараметров; получение итогового списка извлеченных информационных объектов путем выполнения функции «арбитража» конфликтов по отношению к множеству конфликтующих информационных объектов, например, первого информационного объекта из набора конфликтующих информационных объектов, принадлежащего к первому множеству информационных объектов, и второго информационного объекта из набора конфликтующих информационных объектов, принадлежащего ко второму множеству информационных объектов, извлеченных из текста на естественном языке; изменение значений гиперпараметров классификатора извлечения информации; и оптимизация показателя качества извлечения информации для итогового списка извлеченных информационных объектов путем итеративного повторения операций обучения классификатора извлечения информации, выполнения функции «арбитража» конфликтов и изменения значений гиперпараметров.

КРАТКОЕ ОПИСАНИЕ ЧЕРТЕЖЕЙ

[0007] Настоящее изобретение иллюстрируется с помощью примеров, а не способом ограничения и может быть лучше понято при рассмотрении приведенного ниже описания предпочтительных вариантов реализации в сочетании с чертежами, на которых:

[0008] На Фиг. 1 приведена блок-схема одного иллюстративного примера способа извлечения информации с помощью нескольких методов извлечения в соответствии с одним или более вариантами реализации настоящего изобретения;

[0009] На Фиг. 2 схематически показан пример «арбитража» конфликтующих объектов, который может быть выполнен в соответствии с одним или более вариантами реализации настоящего изобретения, в соответствии с одним или более вариантами реализации настоящего изобретения;

[00010] На Фиг. 3 показана блок-схема одного из примеров способа обучения классификатора, используемого для извлечения информации из текстов на естественном языке, в соответствии с одним или более вариантами реализации настоящего изобретения;

[00011] На Фиг. 4 показана блок-схема другого примера способа обучения классификатора, используемого для извлечения информации из текстов на естественном языке, в соответствии с одним или более вариантами реализации настоящего изобретения;

[00012] На Фиг. 5 показана блок-схема одного из иллюстративных примеров способа 400 семантико-синтаксического анализа предложения на естественном языке в соответствии с одним или более вариантами реализации настоящего изобретения;

[00013] На Фиг. 6 схематически показан пример лексико-морфологической структуры предложения в соответствии с одним или более вариантами реализации настоящего изобретения;

[00014] На Фиг. 6 схематически показаны языковые описания, представляющие модель естественного языка, в соответствии с одним или более вариантами реализации настоящего изобретения;

[00015] На Фиг. 8 схематически показаны примеры морфологических описаний в соответствии с одним или более вариантами реализации настоящего изобретения;

[00016] На Фиг. 9 схематически показаны примеры синтаксических описаний в соответствии с одним или более вариантами реализации настоящего изобретения;

[00017] На Фиг. 10 схематически показаны примеры семантических описаний в соответствии с одним или более вариантами реализации настоящего изобретения;

[00018] На Фиг. 11 схематически показаны примеры лексических описаний в соответствии с одним или более вариантами реализации настоящего изобретения;

[00019] На Фиг. 12 схематически показаны примеры структур данных, которые могут быть использованы в рамках одного или более способов, реализованных в соответствии с одним или более вариантами реализации настоящего изобретения;

[00020] На Фиг. 13 схематически показан пример графа обобщенных составляющих в соответствии с одним или более вариантами реализации настоящего изобретения;

[00021] На Фиг. 14 приводится пример синтаксической структуры, соответствующей предложению, приведенному на Фиг. 13;

[00022] На Фиг. 15 показана семантическая структура, соответствующая синтаксической структуре, приведенной на Фиг. 14;

[00023] На Фиг. 16 показана схема примера вычислительной системы, реализующей способы настоящего изобретения.

ОПИСАНИЕ ПРЕДПОЧТИТЕЛЬНЫХ ВАРИАНТОВ РЕАЛИЗАЦИИ

[00024] В настоящем документе описываются способы и системы извлечения информации из текстов на естественном языке. Системы и способы, представленные в настоящем документе, могут найти применение в самых разных ситуациях, где требуется обработка текстов на естественном языке, в частности это могут быть машинный перевод, семантическое индексирование, семантический поиск (в том числе многоязычный семантический поиск), классификация документов, поиск и представление электронных документов (e-discovery) и т.д.

[00025] В настоящем документе термин «вычислительная система» означает устройство обработки данных, оснащенное универсальным или специализированным процессором, памятью и по меньшей мере одним интерфейсом связи. Примерами вычислительных систем, которые могут использовать описанные в этом документе

способы, являются, в частности, настольные компьютеры, ноутбуки, планшетные компьютеры и смартфоны.

[00026] «Онтология» в настоящем документе означает модель, которая представляет объекты, относящиеся к определенной области знаний (предметной области), и отношения между данными объектами. Онтология может включать определения некого множества классов, где каждый класс соответствует концепту предметной области. Каждое определение класса может включать определения одного или более отнесенных к данному классу объектов. Согласно общепринятой терминологии класс онтологии может также называться «концептом», а принадлежащий классу объект может означать экземпляр данного концепта. В некоторых реализациях изобретения класс может быть предком или потомком другого класса. В определенных вариантах реализации объект может быть соотнесен с двумя или более классами.

[00027] Определение каждого класса может далее включать одно или более определений отношений одного или более отнесенных к данному классу объектов. Отношения определяют различные типы взаимодействий между связанными объектами. В некоторых реализациях изобретения различные отношения могут быть организованы во всеобщей таксономии, например, отношения «отцовства» и «материнства» могут быть включены в более общее отношение «быть родителем», которое в свою очередь может быть включено в более общее отношение «быть предком».

[00028] Атрибут информационного объекта может быть представлен другим информационным объектом (например, место рождения человека может быть представлено информационным объектом класса «Город»; а работодатель человека может быть представлен объектом класса «Компания»). Определение класса может включать одно или более ограничений, связанных с определенными атрибутами объекта этого класса, например, атрибут ограничения по кардинальности (например, человек может иметь только одно место рождения) или ограничения по типу (например, заданный атрибут может быть представлен объектами указанного множества классов).

[00029] Извлечение информации может включать анализ текста на естественном языке для выявления и классификации информационных объектов, на которые ссылаются фрагменты текста на естественном языке. Фрагменты текста, ссылающиеся на информационный объект, называются «аннотациями объекта». Фрагменты текста, ссылающиеся на атрибут информационного объекта, называются «аннотациями атрибута». Аннотация может быть определена ее положением в тексте на естественном языке, включая позицию начала (слово) и позицию конца (слово).

[00030] При извлечении информации может происходить классификация выявленных информационных объектов в соответствии с предварительно заданным множеством категорий (например, имена людей, наименования организаций, места, выражения времени, количества, денежных сумм, процентов и т.д.). Такие категории могут быть представлены концептами предварительно заданной или динамически формируемой онтологии. Извлечение информации может дополнительно включать определение отношений между распознанными именованными сущностями и (или) информационными объектами. Примерами таких отношений могут быть работа лица X в организационном подразделении Y, расположение объекта X в геопозиции Y, приобретение организационной единицей X организационной единицы Y и т.д. Подобные отношения могут быть выражены фрагментами на естественном языке, которые могут содержать множество слов из одного или более предложений. Отношения информационных объектов могут быть выражены атрибутами информационного объекта, которые ссылаются на другие информационные объекты (например, информационный объект

«человек» может иметь такие атрибуты, как «место рождения», «место жительства» или «работодатель», каждый из которых представлен ссылкой на информационный объект соответствующего класса).

[00031] Извлеченная информация, которая может быть представлена RDF-графом, может использоваться для выполнения различных операций и задач обработки естественного языка, включая машинный перевод, семантическое индексирование, семантический поиск, классификацию документов, поиск и представление электронных документов (e-discovery) и т.д.

[00032] Извлечение информации может выполняться различными методами извлечения с использованием изменяемых наборов правил, автоматических методов классификации (также называемых «классификаторами на основе машинного обучения»), эвристических подходов и (или) сочетанием этих методов. Методы извлечения информации могут быть основаны на онтологиях и (или) других моделях документов для анализа морфологических, лексических, синтаксических, семантических и (или) иных атрибутов текстов на естественном языке.

[00033] В иллюстративном примере извлечение информации может включать применение множества продукционных правил, например, сопоставление с шаблоном, заданным в левой части продукционного правила, с семантической структурой, представляющей фрагмент текста на естественном языке, может запустить действие правой части продукционного правила, которое может назначить определенные значения одному или более атрибутам информационного объекта. В еще одном иллюстративном примере классификатор на основе машинного обучения может определять степень связанности заданного фрагмента текста на естественном языке (например, представленного набором морфологических, лексических, синтаксических, семантических и (или) иных атрибутов) с определенным классом информационных объектов. Параметры классификатора могут быть определены или изменены в результате дообучения классификатора с использованием ранее существовавших или динамически созданных обучающих выборок данных, на основе которых выполняется корреляция классов информационных объектов с морфологическими, лексическими, синтаксическими, семантическими и (или) иными атрибутами текста на естественном языке. Методы классификации могут включать методы дифференциальной эволюции, генетические алгоритмы, методы случайного леса, нейросети и т.д.

[00034] Различные методы извлечения информации могут иметь различные характеристики производительности, такие как вычислительная точность, полнота и (или) вычислительная сложность. Соответственно, различные методы извлечения информации могут создавать различные результаты при применении к одному и тому же тексту на естественном языке. Определенные информационные объекты, извлеченные на основе двух или более методов извлечения информации, могут конфликтовать друг с другом, например, одно и то же слово на естественном языке может быть распознано как ссылка на географическое положение или имя человека.

[00035] Настоящее изобретение повышает эффективность и качество извлечения информации путем создания систем и способов, которые используют различные сочетания методов извлечения информации с последующим определением и разрешением конфликтов среди извлеченных информационных объектов, таким образом достигается результат, превышающий качество и эффективность традиционно используемых методов. В иллюстративном примере вычислительная система, реализующая этот способ, может применять две или более методов извлечения информации для одного и того же текста на естественном языке. Затем вычислительная система может выявить

множества (например, пары, триплеты, квадруплеты и т.д.) потенциально конфликтующих объектов, извлеченных различными методами, например, путем выявления перекрывающихся текстовых аннотаций и (или) нарушений различных ограничений, связанных с атрибутами информационных объектов (например, атрибут ограничения по кардинальности). После определения множества конфликтующих объектов вычислительная система может предпринять попытку разрешить конфликт за счет применения функции «арбитража» конфликтующих объектов, которая может выполнять анализ морфологических, лексических, синтаксических, семантических и (или) иных атрибутов конфликтующих информационных объектов с целью получения результата, подтверждающего один или более потенциально конфликтующих объектов, изменяющего атрибуты одного или более потенциально конфликтующих объектов, удаляющего один или более потенциально конфликтующих объектов и (или) выполняющего слияние одного или более потенциально конфликтующих объектов.

[00036] В определенных вариантах реализации функция «арбитража» конфликтующих информационных объектов может применять одно или более настраиваемых правил оценки логических условий, определенных для морфологических, лексических, синтаксических, семантических и (или) иных атрибутов извлеченных информационных объектов. Как вариант, функция «арбитража» конфликтующих объектов может включать один или более классификаторов на основе машинного обучения. В иллюстративном примере классификатор, использующийся функцией «арбитража» конфликтующих объектов, может обеспечивать степень обоснованности одного или более возможных результатов «арбитража» (например, слияние возможных конфликтующих объектов в единый объект, подтверждение одного или более извлеченных объектов, изменение одного или более извлеченных объектов и (или) удаление одного или более извлеченных объектов). В другом иллюстративном примере классификатор, использующийся функцией «арбитража» конфликтующих объектов, может обеспечивать сходство двух или более информационных объектов, представляющих один и тот же объект реальной жизни. В еще одном иллюстративном примере классификатор, использующийся функцией «арбитража» конфликтующих объектов, может создавать степени уверенности, связанные с каждым информационным объектом из множества конфликтующих информационных объектов.

[00037] Классификаторы, использующиеся функцией «арбитража» конфликтующих объектов, могут быть обучены на основе существовавших ранее или динамически созданных обучающих выборок данных, на основе которых результаты «арбитража» могут быть соотнесены с морфологическими, лексическими, синтаксическими, семантическими и (или) другими атрибутами извлеченных информационных объектов. Методы классификации могут включать методы дифференциальной эволюции, генетические алгоритмы, методы случайного леса, нейросети и т.д.

[00038] В определенных вариантах реализации обучающие выборки данных для классификаторов на основе машинного обучения могут создаваться путем применения способов на основе правил для отбора текстов на естественном языке и (или) путем верификации пользователем результатов классификации, полученных за счет применения обучаемых классификаторов. В иллюстративном примере классификаторы могут итеративно проходить повторное обучение на основе верифицированных пользователем результатов классификации, за счет чего с каждой итерацией обучения постепенно повышается качество классификации. Контуры такого обучения на основе обратной связи могут быть применены к классификаторам, выполняющим начальное извлечение информационных объектов, а также к классификаторам, реализующим функции

«арбитража» конфликтующих объектов, за счет чего постепенно повышается общее качество извлечения информации.

[00039] В некоторых вариантах осуществления определение классификатора, используемого для извлечения информации, может содержать один или более параметров, значения которых могут регулироваться с помощью методов машинного обучения, и может также включать один или более гиперпараметров, значения которых могут определяться некоторыми внешними по отношению к методам машинного обучения операциями или процессами. В иллюстративном примере значения одного или более гиперпараметров классификатора извлечения информации могут итеративно регулироваться путем итеративного выполнения операций обучения классификатора извлечения информации, обучения классификатора «арбитража» конфликтов, выполнения функции «арбитража» конфликтов и изменения значений гиперпараметров, таким образом итеративно оптимизируя показатель качества, с помощью которого производится оценка итогового списка извлеченных информационных объектов, как более подробно описано ниже в этом документе.

[00040] Системы и способы, работающие в соответствии с одним или более вариантами реализации настоящего изобретения, можно использовать для выполнения различных операций обработки естественного языка, таких как машинный перевод, семантический поиск, классификация и кластеризация объектов и т.д. В некоторых реализациях извлеченная информация может быть визуально представлена с помощью графического пользовательского интерфейса, например, путем отображения идентификаторов классов информационных объектов в визуальном сопоставлении с соответствующими фрагментами текста естественного языка.

[00041] Различные аспекты упомянутых выше способов и систем подробно описаны ниже в этом документе с помощью примеров, а не способом ограничения.

[00042] На Фиг. 1 приведена блок-схема одного иллюстративного примера способа извлечения информации с помощью нескольких методов извлечения в соответствии с одним или более вариантами реализации настоящего изобретения. Способ 100 и (или) каждая из его отдельных функций, стандартных программ, подпрограмм или операций могут выполняться одним или более процессорами вычислительной системы (например, вычислительная система 1000 на Фиг. 16), реализующей этот способ. В некоторых вариантах осуществления способ 100 может выполняться в одном потоке обработки. При альтернативном подходе способ 100 может осуществляться с использованием двух или более потоков обработки, при этом в каждом потоке реализована одна или более отдельных функций, процедур, подпрограмм или действий этого способа. В одном из иллюстративных примеров потоки обработки, в которых реализован способ 100, могут быть синхронизированы (например, с использованием семафоров, критических секций и (или) других механизмов синхронизации потоков). При альтернативном подходе потоки обработки, реализующие способ 100, могут выполняться асинхронно по отношению друг к другу.

[00043] На шаге 110 вычислительная система, реализующая способ, может получать текст на естественном языке. В иллюстративном примере вычислительное устройство может получить текст на естественном языке в виде электронного документа, который может быть получен путем сканирования или за счет применения иного способа изображения с бумажного документа с последующим выполнением оптического распознавания символов (OCR) для получения текста документа. В другом иллюстративном примере вычислительная система может получить текст на естественном языке в виде одного или более форматированных файлов, например,

файлов текстового редактора, сообщений электронной почты, файлов цифровых данных и т.д.

[00044] На шаге 120 вычислительная система может применить к тексту на естественном языке первый метод извлечения информации. В определенных вариантах реализации первый метод извлечения информации может использовать изменяемые наборы правил, автоматические методы классификации (также называемые «классификаторами на основе машинного обучения»), эвристические подходы и (или) сочетание этих способов. При применении первого метода извлечения информации к тексту на естественном языке может создаваться множество информационных объектов одного или более типов информационных объектов.

[00045] В одном из иллюстрационных примеров первый метод извлечения информации может использовать изменяемое множество продукционных правил для интерпретации множества семантико-синтаксических структур, полученных в результате семантико-синтаксического анализа текста на естественном языке. Каждая семантическая структура может представлять предложение на соответствующем естественном языке. Каждая семантическая структура может быть представлена ациклическим графом, который включает множество узлов, соответствующих семантическим классам, и множество дуг, соответствующих семантическим отношениям, что описано более подробно ниже со ссылкой на Фиг. 15.

[00046] Продукционные правила, применяемые ко множеству семантических структур, могут содержать правила интерпретации и правила идентификации. Правило интерпретации может содержать левую часть, представленную множеством логических выражений, определенных на одном или более шаблонах семантической структуры, и правую часть, представленную одним или более утверждениями относительно информационных объектов, представляющих сущности, на которые имеется ссылка в тексте на естественном языке. Шаблон семантической структуры может включать определенные элементы семантической структуры {например, связь с определенным лексическим/семантическим классом, связь с определенной поверхностной или глубинной позицией в указанном месте семантической структуры, наличие отношений предок-потомок между определенными узлами семантической структуры и ряд уровней семантической иерархии между родительскими и дочерними классами в определенной онтологии, наличие общего предка для определенных узлов семантической структуры, наличие определенных граммем или семантем и т.д.). Отношения между элементами семантических структур могут задаваться с помощью одного или более логических выражений (конъюнкция, дизъюнкция и отрицание) и(или) операций, характеризующих взаимное расположение узлов в синтактико-семантическом дереве. В качестве иллюстративного примера такая операция может проверять, принадлежит ли узел к поддереву другого узла.

[00047] Правая часть продукционного правила может назначать и(или) изменять значения одного или более атрибутов информационных объектов, отображающих морфологические, лексические, синтаксические, семантические и(или) иные атрибуты фрагмента текста на естественном языке. В одном из иллюстративных примеров правая сторона правила интерпретации может содержать утверждение, связывающее фрагмент текста на естественном языке с классом информационных объектов.

[00048] Правило идентификации может использоваться для установления ассоциативной связи для пары информационных объектов, которые представляют одну и ту же сущность реального мира. Правило идентификации - это продукционное правило, левая часть которого содержит одно или более логических выражений, указывающих

на узлы семантического дерева, соответствующие информационным объектам. Если указанная пара информационных объектов удовлетворяет условиям, заданным логическими выражениями, то происходит слияние информационных объектов в один информационный объект.

[00049] В иллюстративном примере метод извлечения информации может использовать один или более классификаторов машинного обучения. Каждый классификатор машинного обучения может определять степень связанности заданного фрагмента текста на естественном языке (например, представленного набором морфологических, лексических, синтаксических, семантических и(или) иных атрибутов) с определенным классом информационных объектов. Вычислительная система может итеративно обрабатывать множество фрагментов текста на естественном языке и определять для каждого текстового фрагмента степень его связи с одним или более классами информационных объектов. Методы классификации могут включать методы дифференциальной эволюции, генетические алгоритмы, методы случайного леса, нейросети и т.д.

[00050] Классификаторы могут быть обучены с использованием ранее существовавших и(или) динамически созданных обучающих выборок данных, на основе которых выполняется корреляция классов информационных объектов с лексическими, синтаксическими, семантическими и(или) иными атрибутами текста на естественном языке. Обучающая выборка данных может включать один или более текстов на естественном языке в сопровождении метаданных, которые определяют некоторые информационные объекты, их классификацию, атрибуты и соответствующие текстовые аннотации. В иллюстративном примере метаданные могут быть представлены разметкой, связанной с текстом на естественном языке. В определенных вариантах реализации обучающая выборка данных может итеративно совершенствоваться за счет добавления новых текстов на естественном языке, сопровождаемых подтвержденными пользователем метаданными. Валидация метаданных может включать получение через графический интерфейс пользователя пользовательского ввода, подтверждающего или корректирующего извлеченные информационные объекты и их атрибуты.

[00051] Извлеченные информационные объекты могут быть представлены в виде RDF-графа. RDF (Resource Definition Framework - среда определения ресурса) присваивает каждому информационному объекту уникальный идентификатор и сохраняет информацию о таком объекте в виде наборов из трех элементов (триплетов) SPO, где S означает «субъект» и содержит идентификатор объекта, P означает «предикат» и определяет некоторое свойство этого объекта, а O означает «объект» и хранит в себе значение рассматриваемого свойства данного объекта. Это значение может быть либо примитивным типом данных (примеры: строка, число, булево (логическое) значение), либо идентификатором другого объекта. В одном из иллюстративных примеров триплет SPO может ассоциировать фрагмент из текста на естественном языке с категорией информационных объектов.

[00052] В определенных вариантах реализации методы извлечения информации могут быть основаны на онтологиях и(или) других моделях документов для анализа лексических, синтаксических, семантических и(или) иных атрибутов текста на естественном языке. Таким образом, операциям по извлечению информации, указанным на шаге 120, может предшествовать одна или более операций предварительной обработки документа, выполняемых для определения структуры документа и определения соответствующей модели документа. В одном иллюстративном примере структура документа может включать иерархическую многоуровневую структуру, в

которой разделы документа разделяются заголовками и подзаголовками. В другом иллюстративном примере структура документа может включать одну или более таблиц, содержащих несколько строк и столбцов, по меньшей мере некоторые из которых могут быть связаны с заголовками, которые в свою очередь могут быть организованы в многоуровневую иерархию. В другом иллюстративном примере структура документа может включать структуру страницы, содержащую верхний колонтитул страницы, тело страницы и(или) нижний колонтитул страницы. В другом иллюстративном примере структура документа может включать определенные текстовые поля, связанные с заранее определенными типами информации, такими как поле подписи, поле даты, поле адреса, поле имени и т.д. Вычислительная система 100, в которой реализован этот способ, может интерпретировать структуру документа для получения определенной информации о структуре документа, которая может использоваться для расширения текстовой информации, содержащейся в этом документе. В некоторых реализациях изобретения при анализе структурированных документов вычислительная система может использовать различные вспомогательные онтологии, содержащие классы и концепты, отражающие специфическую структуру документа. Классы вспомогательной онтологии могут быть ассоциированы с определенными продукционными правилами и(или) функциями классификатора, которые могут быть применены к нескольким семантическим структурам, полученным при семантико-синтаксическом анализе соответствующего документа для внесения в результирующее множество семантических структур определенной информации, передаваемой структурой этого документа.

[00053] На шаге 130 вычислительная система может применить к тексту на естественном языке второй метод извлечения информации, который отличается от первого метода извлечения информации. В определенных вариантах реализации второй метод извлечения информации может использовать изменяемые наборы правил, автоматические методы классификации (также называемые «классификаторами машинного обучения»), эвристические подходы и(или) сочетание этих способов, что описано более подробно ниже со ссылкой на шаг 120. При применении второго метода извлечения информации может быть создано второе множество информационных объектов, которое может отличаться от первого множества информационных объектов, созданного при применении первого метода извлечения информации.

[00054] На Фиг. 1 показано только два множества информационных объектов, извлекаемых с применением двух различных методов извлечения информации, однако различные варианты реализации способа 100 могут предусматривать применение двух или более различных методы извлечения информация для извлечения двух или более множеств информационных объектов.

[00055] На шаге 140 вычислительная система может создавать промежуточный 250 перечень информационных объектов, который может включать по меньшей мере подмножество для каждого множества информационных объектов, извлеченных за счет применения различных методов извлечения информации на шагах 120-130. В иллюстративном примере вычислительная система может включать в промежуточный перечень информационные объекты, связанные с одним или более заранее определенными классами информационных объектов, с исключением информационных объектов, связанных с другими классами. В другом иллюстративном примере вычислительная система может исключать из промежуточного перечня информационные объекты одного или более заранее определенных классов информационных объектов. Как описано выше, промежуточный перечень может включать одно или более множеств потенциально конфликтующих информационных объектов, совместное существование

которых нарушает одно или более заданных правил (например, перекрытие текстовых аннотаций или ограничения по кардинальности). Промежуточный перечень может быть представлен RDF-графом, созданным путем слияния RDF-графов, представляющих соответствующие множества извлеченных информационных объектов.

[00056] На шаге 150 вычислительная система может определить один или более множеств конфликтующих информационных объектов, при этом каждое множество включает два или более информационных объекта, по меньшей мере два из которых были извлечены различными методами извлечения информации. В иллюстративном примере множество конфликтующих информационных объектов может включать два информационных объекта, при этом первый информационный объект относится к первому множеству информационных объектов, извлеченных по первому методу извлечения информации, а второй информационный объект относится ко второму множеству информационных объектов, извлеченных по второму методу извлечения информации.

[00057] В иллюстративном примере определение потенциально конфликтующих информационных объектов может включать определение извлеченных информационных объектов, имеющих как минимум частично пересекающиеся текстовые аннотации. Вычислительная система, реализующая этот способ, может итеративно обрабатывать промежуточный перечень извлеченных информационных объектов с целью выявления множеств информационных объектов (например, пары, триплеты, квадруплеты и т.д.) с пересекающимися текстовыми аннотациями. В определенных вариантах реализации вычислительная система может определять, что два или более извлеченных информационных объектов конфликтуют, если их соответствующие аннотации пересекаются по меньшей мере в пороговом количестве слов.

[00058] В другом иллюстративном примере определение потенциально конфликтующих информационных объектов может включать определение различных ограничений, связанных с атрибутами информационных объектов. В определенных вариантах реализации вычислительная система, реализующая этот способ, может выполнять оценку логических условий, определенных для атрибутов информационных объектов (например, сравнение атрибута одного информационного объекта с атрибутом другого информационного объекта). В иллюстративном примере вычислительная система может выявлять нарушение ограничения по кардинальности за счет определения информационного объекта, имеющего ряд атрибутов определенного типа (например, человек может иметь только одну дату рождения и одно место рождения). Кардинальность и другие ограничения могут быть указаны в модели документа или онтологии, связанной с обрабатываемым документом.

[00059] На шаге 160 вычислительная система может применить функцию «арбитража» конфликтующих объектов для каждого выявленного множества конфликтующих информационных объектов для разрешения конфликтов в промежуточном перечне информационных объектов, за счет чего создается окончательный перечень информационных объектов, извлеченных из текстов на естественном языке. Для каждого множества конфликтующих информационных объектов функция «арбитража» конфликтующих объектов может разрешать конфликт путем анализа морфологических, лексических, синтаксических, семантических и(или) иных атрибутов конфликтующих информационных объектов с целью создания результата, подтверждающего один или более потенциально конфликтующих объектов, изменяющего один или более потенциально конфликтующих объектов, удаляющего один или более потенциально конфликтующих объектов и(или) выполняющего слияние одного или более потенциально

конфликтующих объектов. На основе результатов работы функции «арбитража» конфликтующих объектов вычислительная система может изменять RDF-граф, представляющий промежуточный перечень извлеченных информационных объектов, на основе чего создается окончательный RDF-граф.

[00060] На Фиг. 2 схематически показан пример «арбитража» конфликтующих объектов, который может быть выполнен в соответствии с одним или более вариантами реализации настоящего изобретения. Как показано за Фиг. 2, при применении к тексту на естественном языке 210 первого метода извлечения информации 220A создается первое множество информационных объектов 230A, а при применении к тексту на естественном языке 210 второго метода извлечения информации 220B создается второе множество информационных объектов 230B. Первое множество информационных объектов 230A может включать информационный объект 240A, который может быть потенциально конфликтующим с информационным объектом 240 B, входящим во второе множество информационных объектов 230B. При применении функции «арбитража» конфликтующих объектов 250 к потенциально конфликтующим информационным объектам 260 может создаваться один из результатов «арбитража» конфликтующих объектов 260. Результат «арбитража» конфликтующих объектов 260A включает подтверждение информационных объектов 240A и 240B и изменение по меньшей мере некоторых из их атрибутов (например, путем изменения класса информационного объекта и соответствующего разрешения очевидного конфликта). Результат «арбитража» конфликтующих объектов 260B включает удаление информационного объекта 240B. По меньшей мере некоторые из атрибутов удаленного информационного объекта 240B могут быть скопированы в оставшийся информационный объект 240A. Результат «арбитража» конфликтующих объектов 260C включает слияние информационных объектов 240A и 240B в новый информационный объект 240C.

[00061] В определенных вариантах реализации функция «арбитража» конфликтующих информационных объектов может применять одно или более настраиваемых продукционных правил оценки логических условий, определенных для морфологических, лексических, синтаксических, семантических и(или) иных атрибутов извлеченных информационных объектов. В иллюстративном примере реализации атрибуты извлеченных информационных объектов, оцениваемые продукционными правилами, могут включать степени уверенности, связанные с соответствующими информационными объектами. Степени уверенности могут создаваться методами извлечения информации, применяемыми на шагах 120-130.

[00062] Как вариант, функция «арбитража» конфликтующих объектов может включать один или более классификаторов машинного обучения. В иллюстративном примере классификатор, использующийся функцией «арбитража» конфликтующих объектов, может обеспечивать степень обоснованности одного или более возможных результатов «арбитража» (например, слияние возможных конфликтующих объектов в единый объект, подтверждение одного или более извлеченных объектов, изменение одного или более извлеченных объектов и(или) удаление одного или более извлеченных объектов). В другом иллюстративном примере классификатор, использующийся функцией «арбитража» конфликтующих объектов, может фиксировать сходство двух или более информационных объектов, представляющих один и тот же объект реальной жизни. Если вероятность превышает пороговое значение, функция «арбитража» конфликтующих объектов может выполнять слияние выявленных информационных объектов. В еще одном иллюстративном примере классификатор, использующийся

функцией «арбитража» конфликтующих объектов, может создавать степени уверенности, связанные с каждым информационным объектом из множества конфликтующих информационных объектов. Функция «арбитража» конфликтующих объектов может удалять один или более информационных объектов с минимальной степенью уверенности и(или) степенью уверенности ниже заданного порога.

[00063] Классификаторы, использующиеся функцией «арбитража» конфликтующих объектов, могут быть обучены на основе существовавших ранее или динамически созданных обучающих выборок данных, на основе которых результаты «арбитража» могут быть соотнесены с морфологическими, лексическими, синтаксическими, семантическими и(или) другими атрибутами извлеченных информационных объектов. В определенных вариантах реализации атрибуты извлеченных информационных объектов, оцениваемые классификаторами на основе машинного обучения, могут включать степени уверенности, связанные с соответствующими информационными объектами. Методы классификации могут включать методы дифференциальной эволюции, генетические алгоритмы, методы случайного леса, нейросети и т.д.

[00064] Классификатор может быть обучен с использованием ранее существовавших и(или) динамически созданных обучающих выборок данных, на основе которых выполняется корреляция возможных результатов «арбитража» с лексическими, синтаксическими, семантическими и(или) иными атрибутами текста на естественном языке. Обучающая выборка данных может включать один или более текстов на естественном языке в сопровождении метаданных, которые определяют множества конфликтующих информационных объектов, их классификацию, атрибуты и соответствующие результаты «арбитража» конфликтующих объектов. В иллюстративном примере метаданные могут быть представлены разметкой, связанной с текстом на естественном языке. В определенных вариантах реализации обучающая выборка данных может итеративно совершенствоваться за счет добавления новых текстов на естественном языке, сопровождаемых подтвержденными пользователем метаданными. Валидация метаданных может включать получение через графический интерфейс пользователя пользовательского ввода, подтверждающего или корректирующего извлеченные информационные объекты и их атрибуты. В некоторых реализациях верифицированные пользователем тексты и сопровождающие их метаданные, специфицирующие извлеченные информационные объекты, их атрибуты и текстовые аннотации, могут быть использованы для формирования и обновления обучающих выборок, применяемых для обучения классификаторов как подробно описано ниже.

[00065] Со ссылкой на Фиг. 1, на шаге 170 вычислительная система может использовать окончательный перечень информационных объектов для выполнения задач или операций обработки естественного языка, включая машинный перевод, семантическое индексирование, семантический поиск (в том числе семантический поиск на нескольких языках), классификацию документов, поиск и представление электронных документов (e-discovery) и т.д. В некоторых реализациях компьютерная система может отображать извлеченные информационные объекты с визуальными ссылками к фрагментам текста на естественном языке и учитывать действия пользователя, подтверждающие или изменяющие текстовые аннотации и/или результат классификации объекта.

[00066] В некоторых вариантах осуществления определение классификатора, используемого для извлечения информации, может содержать один или более параметров, значения которых могут регулироваться с помощью методов машинного

обучения, и может также включать один или более гиперпараметров, значения которых могут определяться некоторыми внешними по отношению к методам машинного обучения операциями или процессами. Другими словами, значения гиперпараметров классификатора определяются до применения методов машинного обучения для тонкой настройки параметров классификатора.

[00067] В одном из иллюстративных примеров гиперпараметр классификатора извлечения информации может быть представлен параметром регуляризации классификатора градиентного бустинга. В другом иллюстративном примере гиперпараметр классификатора извлечения информации может быть представлен параметром β F-меры, полученным путем оценки информационного объекта, выдаваемого классификатором, который определяется следующим образом:

[00068] $F_\beta=(1+\beta^2) * (\text{Точность} * \text{Полнота}) / ((\beta^2 * \text{Точность}) + \text{Полнота})$,

где $\text{Точность} = t_p / (t_p+f_p)$ и $\text{Полнота} = t_p / (t_p+f_n)$,

t_p - это количество истинно положительных результатов (правильно классифицированных извлеченных информационных объектов), f_p - количество ложно положительных результатов (информационный объект, который не принадлежит к определенному классу, был классифицирован как принадлежащий к этому классу), а f_n -количество ложно отрицательных результатов (информационный объект, который принадлежит к определенному классу, не был классифицирован как принадлежащий к этому классу).

[00069] В некоторых вариантах реализации значения одного или более гиперпараметров классификатора извлечения информации может итеративно регулироваться путем итеративного выполнения операций обучения классификатора извлечения информации, обучения классификатора «арбитража» конфликтов, выполнения функции «арбитража» конфликтов и модификации значений гиперпараметров на основе результата применения функции «арбитража» конфликтов, таким образом производится итеративная оптимизация показателя качества для оценки итогового списка извлеченных информационных объектов, как более подробно будет описано ниже в этом документе со ссылкой на Фиг. 3.

[00070] На Фиг. 3 приведена блок-схема одного примера способа обучения классификатора, используемого для извлечения информации из текстов на естественном языке в соответствии с одним или более вариантами реализации настоящего изобретения. Способ 300 и(или) каждая из его отдельных функций, программ, подпрограмм или операций могут выполняться одним или более процессорами вычислительной системы {например, вычислительной системы 1000 на Фиг. 16), реализующими этот способ. В некоторых вариантах реализации способ 300 может выполняться в одном потоке обработки. При альтернативном подходе способ 300 может быть реализован двумя или более потоками обработки, при этом в каждом потоке реализована одна или несколько отдельных функций, процедур, подпрограмм или операций этого способа. В иллюстративном примере потоки обработки, в которых реализован способ 300, могут быть синхронизированы {например, с использованием семафоров, критических секций и(или) других механизмов синхронизации потоков). При альтернативном подходе потоки обработки, реализующие способ 300, могут выполняться асинхронно по отношению друг к другу.

[00071] На шаге 310 вычислительная система реализует способ, позволяющий обучать классификатор извлечения информации извлечению множества информационных объектов из текста на естественном языке. В некоторых вариантах реализации

определение классификатора может включать один или более параметров, значения которых регулируются при операции обучения классификатора и могут дополнительно включать один или более гиперпараметров, значения которых могут определяться некоторой внешней по отношению к машинному обучению операцией или процессом. В одном из иллюстративных примеров гиперпараметр классификатора извлечения информации может быть представлен параметром регуляризации классификатора градиентного бустинга. В другом иллюстративном примере гиперпараметр классификатора извлечения информации может быть представлен параметром β F-меры, полученным путем оценки информационного объекта, выдаваемого классификатором.

[00072] В одном из иллюстративных примеров аннотированный корпус текстов, используемый для обучения классификатора извлечения информации, может быть разделен на обучающую выборку данных и набор проверочных данных. Обучающая выборка данных затем может быть использована для определения значений параметров классификатора, а набор проверочных данных может использоваться для вычисления показателя качества извлечения информации (то есть, быть представлен параметризованной F-мерой). В одном из иллюстративных примеров процесс обучения может включать настройку одного или более параметров классификатора извлечения информации до тех пор, пока выбранный показатель качества, применяемый к извлеченным информационным объектам, не будет соответствовать установленному заранее пороговому значению.

[00073] Разделение корпуса текстов может включать перекрестную проверку обучающей выборки данных и набора проверочных данных. Для сокращения непостоянства результата может выполняться несколько итераций перекрестной проверки с применением различных разделений, а результаты проверки модели могут быть агрегированы (например, усреднены) по итерациям. В иллюстративном примере к корпусу текстов на естественном языке может применяться способ k-кратной перекрестной проверки. Способ может включать случайное разделение исходного корпуса текстов на k наборов данных одинакового размера, один из которых затем используется в качестве набора проверочных данных, а оставшиеся k-1 дополнительных наборов затем используются как обучающие выборки данных. Затем процесс перекрестной проверки может быть повторен k раз таким образом, чтобы каждый из k наборов данных использовался один раз в качестве проверочных данных. После этого k результатов могут быть агрегированы для создания единого расчета.

[00074] На шаге 320 вычислительная система может использовать классификатор извлечения информации к набору проверочных данных, получая таким образом первое множество информационных объектов.

[00075] На шаге 330 вычислительная система может применять к тому же набору проверочных данных второй метод извлечения информации, который отличается от упомянутого ранее классификатора извлечения информации, таким образом получая второе множество информационных объектов. В одном из иллюстративных примеров второй метод извлечения информации может использовать набор настраиваемых продукционных правил для интерпретации набора семантико-синтаксических структур, полученных путем семантико-синтаксического анализа текста на естественном языке, как более подробно описано ниже в этом документе. В другом иллюстративном примере второй метод извлечения информации может использовать один или более классификаторов машинного обучения, которые отличаются от указанного выше классификатора извлечения информации, который использовался для извлечения

первого множества информационных объектов.

[00076] На шаге 340 вычислительная система может определить один или более множеств конфликтующих информационных объектов, при этом каждое множество включает два или более информационных объекта, по меньшей мере два из которых были извлечены различными методами извлечения информации. В иллюстративном примере множество конфликтующих информационных объектов может включать два информационных объекта, при этом первый информационный объект относится к первому множеству информационных объектов, извлеченных с помощью классификатора извлечения информации, а второй информационный объект относится ко второму множеству информационных объектов, извлеченных по другому методу извлечения информации, как более подробно было рассмотрено выше.

[00077] На шаге 350 вычислительная система может применить функцию арбитража конфликтов к каждому обнаруженному множеству конфликтующих информационных объектов, чтобы разрешить конфликты, таким образом, создается окончательный список информационных объектов, извлеченных из текстов на естественном языке, как описано более подробно выше.

[00078] В иллюстративном примере функция «арбитража» конфликтующих объектов, может быть реализована классификатором на основе машинного обучения, обученным производить степень "обоснованности" одного или более возможных результатов «арбитража» (например, слияние возможных конфликтующих объектов в единый объект, подтверждение одного или более извлеченных объектов, изменение одного или более извлеченных объектов и(или) удаление одного или более извлеченных объектов). В другом иллюстративном примере классификатор, использующийся функцией «арбитража» конфликтующих объектов, может обеспечивать сходство двух или более информационных объектов, представляющих один и тот же объект реальной жизни. Если вероятность превышает пороговое значение, функция «арбитража» конфликтующих объектов может выполнять слияние выявленных информационных объектов. В другом иллюстративном примере классификатор, использующийся функцией «арбитража» конфликтующих объектов, может создавать степени уверенности, связанные с каждым информационным объектом из множества конфликтующих информационных объектов. Функция «арбитража» конфликтующих объектов может удалять один или более информационных объектов с минимальной степенью уверенности и(или) степенью уверенности ниже заданного порога.

[00079] Альтернативно, функция арбитража конфликтов может быть реализована одним или несколькими настраиваемыми правилами, оценивающими логические условия, определенные на морфологических, лексических, синтаксических, семантических и/или других атрибутах извлеченных информационных объектов, как описано более подробно здесь.

[00080] На шаге 360 вычислительная система может оценивать заранее определенный показатель качества, применяемый к итоговому набору извлеченных информационных объектов. В одном из иллюстративных примеров показатель качества обеспечивается при помощи F-меры.

[00081] В ответ на определение на шаге 370 того, что заранее определенные условия прекращения не выполнены, обработка продолжается на шаге 380 изменением одного или более гиперпараметров классификатора извлечения информации. В одном из иллюстративных примеров гиперпараметр может быть представлен параметром β F-меры, который используется для оценки качества извлечения информации. Таким образом, вычислительная система может модифицировать значение параметра β для

повышения точности или полноты результатов, получаемых от классификатора извлечения информации. В некоторых вариантах осуществления вычислительная система может реализовывать поиск по сетке параметров, итеративно изменяя значения одного или более гиперпараметров на заранее определенные приращения так, чтобы параметры оставались внутри определенных диапазонов, таким образом оптимизируя показатель качества, применяемый для итогового набора извлеченных информационных объектов. После выполнения операций на шаге 380 способ может вернуться к шагу 310.

[00082] В ответ на определение на шаге 370 факта выполнения условий прекращения способ может выйти из цикла и продолжить работу на шаге 390. В одном из иллюстративных примеров условия прекращения могут определяться по максимальному количеству выполненных итераций. В ответ на определение факта выполнения указанного числа операций вычислительная система может на шаге 390 выявить среди значений, указанных выше, заранее определенных показателей качества, оцененных на шаге 370, которые создавались на всех выполненных итерациях, оптимальное (то есть минимальное или максимальное) значение показателя качества и соответствующим образом выявить значения гиперпараметров классификатора извлечения информации, который использовался в итерации, которая показала оптимальный показатель качества.

[00083] В другом иллюстративном примере условия прекращения, проверенные на шаге 370, могут определять пороговое значение, соответствующее показателю качества, который оценивался на шаге 370. В ответ на определение факта соответствия показателя качества указанному пороговому значению вычислительная система может на шаге 390 выявить значения гиперпараметров классификатора извлечения информации, который был использован в последней выполненной итерации, которая дала значение показателя качества, соответствующее условиям прекращения выполнения. В ответ на определение значений гиперпараметров классификатора извлечения информации способ может прекратить работу. Обученный классификатор извлечения информации может быть затем использован в способе 100 извлечения информации.

[00084] Как отмечено выше, в некоторых реализациях функция арбитража конфликтов может выполняться посредством классификатора на основе машинного обучения. Определение такого классификатора может включать в себя один или несколько параметров, значения которых могут настраиваться с помощью методов машинного обучения и могут дополнительно включать в себя один или несколько гиперпараметров, значения которых могут быть определены некоторыми внешними по отношению к машинному обучению, операциями или процессом. Другими словами, значения гиперпараметров классификатора определяются до применения методов машинного обучения для точной настройки параметров классификатора.

[00085] В иллюстративном примере гиперпараметр классификатора арбитража конфликтов может быть представлен параметром регуляризации классификатора градиентного бустинга. В другом иллюстративном примере гиперпараметр классификатора арбитража конфликтов может быть представлен параметром β показателя F-меры путем оценки информационных объектов, полученных классификатором.

[00086] В некоторых реализациях значения одного или нескольких гиперпараметров классификатора арбитража конфликтов могут итеративно корректироваться путем итеративного выполнения операций извлечения первого множества информационных объектов, извлечения второго множества информационных объектов, изменения значений гипер-параметров на основе оценки качества извлечения информации, и обучение классификатора арбитража конфликтов, таким образом итеративно

оптимизируя метрику качества, оценивая окончательный список извлеченных информационных объектов. В некоторых реализациях компьютерная система может реализовать поиск лучших параметров по сетке (grid search) путем итеративного изменения значений одного или нескольких гиперпараметров с помощью заранее определенных приращений, в то время как параметры остаются в определенных диапазонах, таким образом оптимизируя метрику качества, применяемую к окончательному набору извлеченных информационных объектов, как более подробно описано ниже со ссылкой на Фиг. 4.

[00087] На Фиг. 4 показана блок-схема другого примера способа обучения классификатора, используемого для извлечения информации из текстов на естественном языке, в соответствии с одним или несколькими аспектами настоящего раскрытия. Способ 400 и/или каждая из его отдельных функций, подпрограмм, подпрограмм или операций может выполняться одним или несколькими процессорами компьютерной системы (например, компьютерной системой 1000 на Фиг. 16), реализующей этот метод. В некоторых реализациях способ 400 может выполняться одним потоком обработки. Альтернативно, способ 400 может выполняться двумя или более потоками обработки, причем каждый поток реализует одну или несколько отдельных функций, подпрограмм, подпрограмм или операций этого метода. В иллюстративном примере способ реализации 400 потоков обработки может быть синхронизирован (например, с использованием семафоров, критических секций и/или других механизмов синхронизации потоков). Альтернативно, способ 400 реализации потоков обработки может выполняться асинхронно относительно друг друга.

[00088] На этапе 410 компьютерная система, реализующая способ, может применять к тексту на естественном языке первый метод извлечения информации. В некоторых реализациях первый метод извлечения информации может использовать настраиваемые наборы правил, автоматические методы классификации (также известные как «классификаторы на основе машинного обучения»), эвристические подходы и/или их комбинации. Применение первого метода извлечения информации к тексту на естественном языке может создавать первое множество информационных объектов одного или нескольких типов информационных объектов.

[00089] На этапе 420 компьютерная система может применять к тексту на естественном языке второй метод извлечения информации, которая отличается от первого метода извлечения информации. В некоторых реализациях второй метод извлечения информации может использовать настраиваемые наборы правил, автоматические методы классификации (также известные как «классификаторы на основе машинного обучения»), эвристические подходы и/или их комбинации. Применение второго метода извлечения информации к тексту на естественном языке может создавать второе множество информационных объектов, которое может отличаться от первого множества информационных объектов, созданных с помощью первого метода извлечения информации.

[00090] На этапе 430 компьютерная система может изменять значения одного или нескольких гиперпараметров классификатора арбитража конфликтов. В иллюстративном примере гиперпараметр может быть представлен параметром регуляризации метода градиентного бустинга. В другом иллюстративном примере гиперпараметр может быть представлен параметром β метрики F-меры путем оценки информационных объектов, полученных классификатором. В некоторых реализациях компьютерная система может реализовать поиск лучших параметров по сетке (grid search) путем итеративного изменения значений одного или нескольких гиперпараметров

с помощью заранее определенных приращений, в то время как параметры остаются в определенных диапазонах, таким образом оптимизируя метрику качества, применяемую к окончательному набору извлеченных информационных объектов.

[00091] На этапе 440 компьютерная система может обучать классификатор арбитража конфликтов для выполнения функции арбитража конфликтов в отношении набора конфликтующих информационных объектов, чтобы составить окончательный список извлеченных информационных объектов. В иллюстративном примере набор конфликтующих информационных объектов может содержать два информационных объекта, так что первый информационный объект принадлежит первому множеству информационных объектов, который был извлечен классификатором извлечения информации, тогда как второй информационный объект принадлежит второму множеству информационных объектов, которые были извлечены с помощью второго метода извлечения информации, как описано более подробно выше.

[00092] В иллюстративном примере классификатор, используемый функцией арбитража конфликтов, может выдать степень обоснованности одного или нескольких возможных результатов арбитража (например, объединение двух или более потенциально конфликтующих объектов в один объект, подтверждение одного или нескольких извлеченных объектов, изменение одного или нескольких извлеченных объектов и/или удаление одного или нескольких извлеченных объектов). В другом иллюстративном примере классификатор, используемый функцией арбитража конфликтов, может оценивать вероятность того, что два или более информационных объекта представляют один и тот же объект реальной жизни. Если вероятность превышает пороговое значение, функция арбитража конфликтов может объединять идентифицированные информационные объекты. В другом иллюстративном примере классификатор, используемый функцией арбитража конфликтов, может продуцировать степени достоверности, связанные с каждым информационным объектом из множества конфликтующих информационных объектов. Функция арбитража конфликтов может удалить один или несколько информационных объектов, имеющих наименьшие степени достоверности и/или степени достоверности, которые находятся ниже заранее определенного порога.

[00093] В некоторых реализациях определение классификатора арбитража конфликтов может включать один или несколько параметров, значения которых настраиваются с помощью процедуры обучения классификатора и могут дополнительно включать один или несколько гиперпараметров, значения которых могут быть определенных некоторым внешним, по отношению к процессу машинного обучения, процессом или операцией.

[00094] Методы классификации могут включать методы дифференциальной эволюции, генетические алгоритмы, методы случайного леса, нейронные сети и т.д. В иллюстративном примере процесс обучения может включать настройку одного или нескольких параметров классификатора извлечения информации до тех пор, пока не будет достигнут предопределенный порог выбранной метрики качества.

[00095] Классификаторы, используемые в функции арбитража конфликтов, могут быть обучены с использованием ранее существовавших и/или динамически созданных обучающих наборов данных, которые устанавливают соотношение результатов арбитража с морфологическими, лексическими, синтаксическими, семантическими и/или другими атрибутами извлеченных информационных объектов.

[00096] Обучающий набор данных может содержать один или несколько текстов на естественном языке, сопровождаемых метаданными, которые задают множество

конфликтующих информационных объектов, их классификацию, атрибуты и соответствующие результаты арбитража. В иллюстративном примере метаданные могут быть представлены разметкой, связанной с текстом на естественном языке. В некоторых реализациях обучающий набор данных может быть итеративно расширен путем добавления новых текстов на естественном языке, сопровождаемых проверенными пользователем метаданными. Валидация метаданных может включать в себя получение через графический интерфейс пользователя (GUI), пользовательского ввода, подтверждающего или корректирующего извлеченные информационные объекты и их атрибуты.

[00097] В иллюстративном примере аннотированный текстовый корпус, используемый для обучения классификатора арбитража конфликтов, может быть разделен на обучающий набор данных и проверочный набор данных. Затем обучающий набор данных может быть использован для определения значений параметров классификатора, тогда как проверочный набор данных может использоваться для вычисления метрики качества извлечения информации (например, представленной параметризованной F-мерой). В иллюстративном примере процесс обучения может включать в себя корректировку одного или нескольких параметров классификатора арбитража конфликтов до тех пор, пока выбранная метрика качества, применяемая к окончательному списку извлеченных информационных объектов, не будет соответствовать заранее установленному порогу.

[00098] На этапе 450 компьютерная система может применять функцию арбитража конфликта для каждого обнаруженного набора конфликтующих информационных объектов для разрешения конфликтов, создавая таким образом окончательный список информационных объектов, извлеченных из текстов на естественном языке, как описано более подробно выше.

[00099] На этапе 460 компьютерная система может вычислять предопределенную метрику качества применительно к окончательному набору извлеченных информационных объектов. В иллюстративном примере метрика качества может быть представлена посредством F-меры.

[000100] В ответ на установление на этапе 470 того, что предопределенное условие завершения не выполнено, способ может возвращаться к блоку 410; в противном случае способ может выйти из цикла и продолжить обработку в блоке 480. В иллюстративном примере условие завершения может указывать максимальное количество итераций, которые должны быть выполнены. В ответ на выполнение указанного количества итераций компьютерная система может на этапе 470 идентифицировать среди значений вышеописанной предопределенной метрики качества, вычисленных на этапе 460, которые были получены при всех выполненных итерациях, оптимальное (например, максимальное или минимальное) значение метрики качества и, соответственно, может идентифицировать те значения гиперпараметров классификатора арбитража конфликтов, использованные в процессе итерации, которые дали оптимальную метрику качества.

[000101] В другом иллюстративном примере условие завершения, которое оценивается в блоке 470, может указывать пороговое значение, которое должно удовлетворяться метрикой качества, которая была вычислена на этапе 460. В ответ на определение того, что метрика качества соответствует указанному пороговому значению, компьютерная система может на этапе 490 идентифицировать значения гиперпараметров классификатора извлечения информации, которые использовались на последней выполненной итерации, что дало показатель качества, удовлетворяющий условию завершения. В ответ на выявление значений гиперпараметров классификатора арбитража

конфликтов метод может завершиться. Затем обученный классификатор арбитража конфликтов может быть использован методом 100 для извлечения информации.

[000102] Как указано выше, продукционные правила, классификаторы и(или) иные методы извлечения информации и «арбитража» конфликтующих объектов могут выполнять анализ различных морфологических, лексических, синтаксических, семантических и(или) иных атрибутов текста на естественном языке. Такие атрибуты могут создаваться путем выполнения семантико-синтаксического анализа текста на естественном языке, что схематически показано на Фиг. 5.

[000103] На Фиг. 5 приведена блок-схема одного иллюстративного примера реализации способа 200 для выполнения семантико-синтаксического анализа предложения на естественном языке 212 в соответствии с одним или несколькими аспектами настоящего изобретения. Способ 400 может быть применен к одной или более синтаксическим единицам (например, предложениям), включенным в определенный текстовый корпус, для формирования множества семантико-синтаксических деревьев, соответствующих синтаксическим единицам. В различных иллюстративных примерах подлежащие обработке способом 400 предложения на естественном языке могут извлекаться из одного или нескольких электронных документов, которые могут создаваться путем сканирования (или другим способом получения изображений бумажных документов) и оптического распознавания символов (OCR) для получения текстов, соответствующих этим документам. Предложения на естественном языке также могут извлекаться из других различных источников, включая сообщения, отправляемые по электронной почте, тексты из социальных сетей, файлы с цифровым содержимым, обработанные с использованием способов распознавания речи и т.д.

[000104] В блоке 214 вычислительное устройство, реализующее данный способ, может проводить лексико-морфологический анализ предложения 212 для установления морфологических значений слов, входящих в состав предложения. В настоящем документе "морфологическое значение" слова означает одну или несколько лемм (т.е. канонических или словарных форм), соответствующих слову, и соответствующий набор значений грамматических признаков, которые определяют грамматическое значение слова. В число таких грамматических признаков могут входить лексическая категория (часть речи) слова и один или более морфологических и грамматических признаков (например, падеж, род, число, спряжение и т.д.). Ввиду омонимии и(или) совпадающих грамматических форм, соответствующих разным лексико-морфологическим значениям определенного слова, для данного слова может быть установлено два или более морфологических значений. Более подробное описание иллюстративного примера проведения лексико-морфологического анализа предложения более детально приведено ниже в настоящем документе со ссылкой на Фиг. 6.

[000105] В блоке 215 вычислительное устройство может проводить грубый синтаксический анализ предложения 212. Грубый синтаксический анализ может включать применение одной или нескольких синтаксических моделей, которые могут быть соотнесены с элементами предложения 212, с последующим установлением поверхностных (т.е. синтаксических) связей в рамках предложения 212 для получения графа обобщенных составляющих. В настоящем документе "составляющая" означает группу соседних слов исходного предложения, функционирующую как одна грамматическая сущность. Составляющая включает в себя ядро в виде одного или более слов и может также включать одну или несколько дочерних составляющих на более низких уровнях. Дочерняя составляющая является зависимой составляющей,

которая может быть соотнесена с одной или несколькими родительскими составляющими.

[000106] В блоке 216 вычислительное устройство может проводить точный синтаксический анализ предложения 212 для формирования одного или более синтаксических деревьев предложения. Среди различных синтаксических деревьев на основе определенной функции оценки с учетом совместимости лексических значений слов исходного предложения, поверхностных отношений, глубинных отношений и т.д. может быть отобрано одно или несколько лучших синтаксических деревьев, соответствующих предложению 212.

[000107] В блоке 217 вычислительное устройство может обрабатывать синтаксические деревья для формирования семантической структуры 218, соответствующей предложению 212. Семантическая структура 218 может включать множество узлов, соответствующих семантическим классам и также может включать множество дуг, соответствующих семантическим отношениям (более подробное описание см. ниже в настоящем документе).

[000108] Фиг. 6 схематически иллюстрирует пример лексико-морфологической структуры предложения в соответствии с одним или более аспектами настоящего изобретения. Пример лексико-морфологической структуры 300 может включать множество пар "лексическое значение - грамматическое значение" для примера предложения. В качестве иллюстративного примера, "ll" может быть соотнесено с лексическим значением "shall" 512 и "will" 514. Грамматическим значением, соотнесенным с лексическим значением 512, является<Verb, GTVerbModal, ZeroType, Present, Nonnegative, Composite II>. Грамматическим значением, соотнесенным с лексическим значением 314, является<Verb, GTVerbModal, ZeroType, Present, Nonnegative, Irregular, Composite II>.

[000109] Фиг. 7 схематически иллюстрирует используемые языковые описания 210, в том числе морфологические описания 201, лексические описания 203, синтаксические описания 202 и семантические описания 204, а также отношения между ними. Среди них морфологические описания 201, лексические описания 203 и синтаксические описания 202 зависят от языка. Набор языковых описаний 610 представляет собой модель определенного естественного языка.

[000110] В качестве иллюстративного примера определенное лексическое значение в лексических описаниях 203 может быть соотнесено с одной или несколькими поверхностными моделями синтаксических описаний 202, соответствующих данному лексическому значению. Определенная поверхностная модель синтаксических описаний 202 может быть соотнесена с глубинной моделью семантических описаний 204.

[000111] На Фиг. 8 схематически иллюстрируются несколько примеров морфологических описаний. В число компонентов морфологических описаний 201 могут входить: описания словоизменения 710, грамматическая система 720, описания словообразования 730 и другие. Грамматическая система 720 включает набор грамматических категорий, таких как часть речи, падеж, род, число, лицо, возвратность, время, вид и их значения (так называемые "граммемы"), в том числе, например, прилагательное, существительное или глагол; именительный, винительный или родительный падеж; женский, мужской или средний род и т.д. Соответствующие граммемы могут использоваться для составления описания словоизменения 710 и описания словообразования 730.

[000112] Описание словоизменения 710 определяет формы данного слова в зависимости от его грамматических категорий (например, падеж, род, число, время и т.д.) и в широком смысле включает в себя или описывает различные возможные формы

слова. Описание словообразования 730 определяет, какие новые слова могут быть образованы от данного слова (например, сложные слова).

[000113] В соответствии с одним из аспектов настоящего изобретения при установлении синтаксических отношений между элементами исходного предложения могут использоваться модели составляющих. Составляющая представляет собой группу соседних слов в предложении, ведущих себя как единое целое. Ядром составляющей является слово, она также может содержать дочерние составляющие более низких уровней. Дочерняя составляющая является зависимой составляющей и может быть прикреплена к другим составляющим (родительским) для построения синтаксических описаний 202 исходного предложения.

[000114] На Фиг. 9 приведены примеры синтаксических описаний. В число компонентов синтаксических описаний 202 могут входить, среди прочего, поверхностные модели 910, описания поверхностных позиций 920, описание референциального и структурного контроля 956, описание управления и согласования 940, описание недревесного синтаксиса 950 и правила анализа 960. Синтаксические описания 202 могут использоваться для построения возможных синтаксических структур исходного предложения на заданном естественном языке с учетом свободного линейного порядка слов, недревесных синтаксических явлений (например, согласование, эллипсис и т.д.), референциальных отношений и других факторов.

[000115] Поверхностные модели 910 могут быть представлены в виде совокупностей одной или нескольких синтаксических форм («синтформ» 912) для описания возможных синтаксических структур предложений, входящих в состав синтаксического описания 202. В целом, лексическое значение слова на естественном языке может быть связано с поверхностными (синтаксическими) моделями 910. Поверхностная модель может представлять собой составляющие, которые возможны, если лексическое значение выступает в роли "ядра". Поверхностная модель может включать набор поверхностных позиций дочерних элементов, описание линейного порядка и(или) диатезу. В настоящем документе "диатеза" означает определенное отношение между поверхностными и глубинными позициями и их семантическими ролями, выражаемыми посредством глубинных позиций. Например, диатеза может быть выражаться залогом глагола: если субъект является агентом действия, глагол в активном залоге, а когда субъект является направлением действия, это выражается пассивным залогом глагола.

[000116] В модели составляющих может использоваться множество поверхностных позиций 915 дочерних составляющих и описаний их линейного порядка 916 для описания грамматических значений 914 возможных заполнителей этих поверхностных позиций. Диатезы 917 представляют собой соответствия между поверхностными позициями 915 и глубинными позициями 514 (как показано на Фиг. 11). Коммуникативные описания 980 описывают коммуникативный порядок в предложении.

[000117] Описание линейного порядка (916) может быть представлено в виде выражений линейного порядка, отражающих последовательность, в которой различные поверхностные позиции (915) могут встречаться в предложении. В число выражений линейного порядка могут входить наименования переменных, имена поверхностных позиций, круглые скобки, граммемы, оператор «or» (или) и т.д. В качестве иллюстративного примера описание линейного порядка простого предложения "Boys play football" можно представить в виде "Subject Core Object_Direct" (Подлежащее - Ядро - Прямое дополнение), где Subject (Подлежащее), Core (Ядро) и Object_Direct (Прямое дополнение) представляют собой имена поверхностных позиций 915, соответствующих порядку слов.

[000118] Коммуникативные описания 980 могут описывать порядок слов в синтформе 912 с точки зрения коммуникативных актов, представленных в виде коммуникативных выражений порядка, которые похожи на выражения линейного порядка. Описания управления и согласования 940 может включать правила и ограничения на грамматические значения присоединяемых составляющих, которые используются во время синтаксического анализа.

[000119] Описания не древесного синтаксиса 950 могут создаваться для отражения различных языковых явлений, таких как эллипсис и согласование, они используются при трансформациях синтаксических структур, которые создаются на различных этапах анализа в различных вариантах реализации изобретения. Описания недревесного синтаксиса 950 могут, среди прочего, включать описание эллипсиса 952, описания согласования 954, а также описания референциального и структурного контроля 930.

[000120] Правила анализа 960 могут описывать свойства конкретного языка и использоваться в рамках семантического анализа. Правила анализа 960 могут включать правила вычисления семантем 962 и правила нормализации 964. Правила нормализации 964 могут использоваться для описания трансформаций семантических структур, которые могут отличаться в разных языках.

[000121] На Фиг. 10 приведен пример семантических описаний. Компоненты семантических описаний 204 не зависят от языка и могут, среди прочего, включать семантическую иерархию 510, описания глубинных позиций 520, систему семантем 530 и прагматические описания 540.

[000122] Ядро семантических описаний может быть представлено семантической иерархией 510, в которую могут входить семантические понятия (семантические сущности), также называемые семантическими классами. Последние могут быть упорядочены в иерархическую структуру, отражающую отношения "родитель-потомок".

В целом, дочерний семантический класс может унаследовать одно или более свойств своего прямого родителя и других семантических классов-предков. В качестве иллюстративного примера семантический класс SUBSTANCE (Вещество) является дочерним семантическим классом класса ENTITY (Сущность) и родительским семантическим классом для классов GAS, (Газ), LIQUID (Жидкость), METAL (Металл), WOOD MATERIAL (Древесина) и т.д.

[000123] Каждый семантический класс в семантической иерархии 510 может сопровождаться глубинной моделью 512. Глубинная модель 512 семантического класса может включать множество глубинных позиций 514, которые могут отражать семантические роли дочерних составляющих в различных предложениях с объектами данного семантического класса в качестве ядра родительской составляющей. Глубинная модель 512 также может включать возможные семантические классы, выступающие в роли заполнителей глубинных позиций. Глубинные позиции (514) могут выражать семантические отношения, в том числе, например, "agent" (агенс), "addressee" (адресат), "instrument" (инструмент), "quantity" (количество) и т.д. Дочерний семантический класс может наследовать и уточнять глубинную модель своего непосредственного родительского семантического класса.

[000124] Описания глубинных позиций 520 отражают семантические роли дочерних составляющих в глубинных моделях 512 и могут использоваться для описания общих свойств глубинных позиций 514. Описания глубинных позиций 520 также могут содержать грамматические и семантические ограничения в отношении заполнителей глубинных позиций 514. Свойства и ограничения, связанные с глубинными позициями 514 и их возможными заполнителями в различных языках, могут быть в значительной

степени подобными и зачастую идентичными. Таким образом, глубинные позиции 514 не зависят от языка.

[000125] Система семантем 530 может представлять собой множество семантических категорий и семантем, которые представляют значения семантических категорий. В качестве иллюстративного примера семантическая категория "DegreeOflComparison" (Степень сравнения) может использоваться для описания степени сравнения прилагательных и включать следующие семантемы: "Positive" (Положительная), "ComparativeHigherDegree" (Сравнительная степень сравнения), "SuperlativeHighestDegree" (Превосходная степень сравнения) и другие. В качестве еще одного иллюстративного примера семантическая категория "RelationToReferencePoint" (Отношение к точке) может использоваться для описания порядка (пространственного или временного в широком смысле анализируемых слов), как, например, до или после точки или события, и включать семантемы "Previous" (Предыдущий) и "Subsequent" (Последующий). В качестве еще одного иллюстративного примера семантическая категория "EvaluationObjective" (Оценка) может использоваться для описания объективной оценки, как, например, "Bad" (Плохой), "Good" (Хороший) и т.д.

[000126] Система семантем 530 может включать независимые от языка семантические атрибуты, которые могут выражать не только семантические характеристики, но и стилистические, прагматические и коммуникативные характеристики. Некоторые семантемы могут использоваться для выражения атомарного значения, которое находит регулярное грамматическое и(или) лексическое выражение в естественном языке. По своему целевому назначению и использованию системы семантем могут разделяться на категории, например, грамматические семантемы 532, лексические семантемы 534 и классифицирующие грамматические (дифференцирующие) семантемы 536.

[000127] Грамматические семантемы 532 могут использоваться для описания грамматических свойств составляющих при преобразовании синтаксического дерева в семантическую структуру. Лексические семантемы 534 могут описывать конкретные свойства объектов (например, "being flat" (быть плоским) или "being liquid" (являться жидкостью)) и использоваться в описаниях глубинных позиций 520 как ограничение заполнителей глубинных позиций (например, для глаголов "face (with)" (облицовывать) и "flood" (заливать), соответственно). Классифицирующие грамматические (дифференцирующие) семантемы 536 могут выражать дифференциальные свойства объектов внутри одного семантического класса. В качестве иллюстративного примера в семантическом классе HAIRDRESSER (ПАРИКМАХЕР) семантема «RelatedToMen» (Относится к мужчинам) присваивается лексическому значению "barber" в отличие от других лексических значений, которые также относятся к этому классу, например, «hairdresser», «hairstylist» и т.д. Используя данные независимые от языка семантические свойства, которые могут быть выражены в виде элементов семантического описания, в том числе семантических классов, глубинных позиций и семантем, можно извлекать семантическую информацию в соответствии с одним или более аспектами настоящего изобретения.

[000128] Прагматические описания 540 позволяют назначать определенную тему, стиль или жанр текстам и объектам семантической иерархии 510 (например, «Экономическая политика», «Внешняя политика», «Юриспруденция», «Законодательство», «Торговля», «Финансы» и т.д.). Прагматические свойства также могут выражаться семантемами. В качестве иллюстративного примера прагматический контекст может приниматься во внимание при семантическом анализе.

[000129] На Фиг. 11 приведен пример лексических описаний. Лексические описания

(203) представляют собой множество лексических значений 612 конкретного естественного языка. Для каждого лексического значения 612 имеется связь 602 с его независимым от языка семантическим родителем для того, чтобы указать положение какого-либо заданного лексического значения в семантической иерархии 510.

[000130] Лексическое значение 612 в лексико-семантической иерархии 510 может быть соотнесено с поверхностной моделью 910, которая в свою очередь через одну или несколько диатез 917 может быть соотнесена с соответствующей глубинной моделью 512. Лексическое значение 612 может наследовать семантический класс своего родителя и уточнять свою глубинную модель 512.

[000131] Поверхностная модель 910 лексического значения может включать одну или несколько синтаксических форм 912. Синтформа 912 поверхностной модели 910 может включать одну или несколько поверхностных позиций 915, в том числе соответствующие описания их линейного порядка 916, одно или несколько грамматических значений 914, выраженных в виде набора грамматических категорий (граммем), одно или несколько семантических ограничений, соотнесенных с заполнителями поверхностных позиций, и одну или несколько диатез 917. Семантические ограничения, соотнесенные с определенным заполнителем поверхностной позиции, могут быть представлены в виде одного или более семантических классов, объекты которых могут заполнить эту поверхностную позицию.

[000132] На Фиг. 12 схематически иллюстрируются примеры структур данных, которые могут быть использованы в рамках одного или более методов настоящего изобретения. Снова ссылаясь на Фиг. 5, в блоке 214 вычислительное устройство, реализующее данный способ, может проводить лексико-морфологический анализ предложения 212 для построения лексико-морфологической структуры 722 согласно Фиг. 12. Лексико-морфологическая структура 722 может включать множество соответствий лексического и грамматического значений для каждой лексической единицы (например, слова) исходного предложения. Фиг. 6 схематически иллюстрирует пример лексико-морфологической структуры.

[000133] Снова возвращаясь к Фиг. 6, в блоке 215 вычислительное устройство может проводить грубый синтаксический анализ исходного предложения 212 для построения графа обобщенных составляющих 732 согласно Фиг. 12. Грубый синтаксический анализ предполагает применение одной или нескольких возможных синтаксических моделей возможных лексических значений к каждому элементу множества элементов лексико-морфологической структуры 722, с тем чтобы установить множество потенциальных синтаксических отношений в составе исходного предложения 212, представленных графом обобщенных составляющих 732.

[000134] Граф обобщенных составляющих 732 может быть представлен ациклическим графом, включающим множество узлов, соответствующих обобщенным составляющим исходного предложения 212 и включающим множество дуг, соответствующих поверхностным (синтаксическим) позициям, которые могут выражать различные типы отношений между обобщенными лексическими значениями. В рамках данного способа может применяться множество потенциально применимых синтаксических моделей для каждого элемента множества элементов лексико-морфологических структур исходного предложения 212 для формирования набора составляющих исходного предложения 212. Затем в рамках способа может рассматриваться множество возможных составляющих исходного предложения 212 для построения графа обобщенных составляющих 732 на основе набора составляющих. Граф обобщенных составляющих 732 на уровне поверхностной модели может отражать множество потенциальных связей

между словами исходного предложения 212. Поскольку количество возможных синтаксических структур может быть относительно большим, граф обобщенных составляющих 732 может, в общем случае, включать избыточную информацию, в том числе относительно большое число лексических значений по определенным узлам и (или) поверхностных позиций по определенным дугам графа.

[000135] Граф обобщенных составляющих 732 может изначально строиться в виде дерева, начиная с концевых узлов (листьев) и двигаясь далее к корню, путем добавления дочерних составляющих, заполняющих поверхностные позиции 915 множества родительских составляющих, с тем чтобы были охвачены все лексические единицы исходного предложения 212.

[000136] В некоторых вариантах осуществления корень графа обобщенных составляющих 732 представляет собой предикат. В ходе описанного выше процесса дерево может стать графом, так как определенные составляющие более низкого уровня могут быть включены в одну или несколько составляющих верхнего уровня. Множество составляющих, которые представляют определенные элементы лексико-морфологической структуры, затем может быть обобщено для получения обобщенных составляющих. Составляющие могут быть обобщены на основе их лексических значений или грамматических значений 914, например, на основе частей речи и отношений между ними. Фиг. 13 схематически иллюстрирует пример графа обобщенных составляющих.

[000137] В блоке 216 вычислительное устройство может проводить точный синтаксический анализ предложения 212 для формирования одного или более синтаксических деревьев 742 согласно Фиг. 12 на основе графа обобщенных составляющих 732. Для каждого синтаксического дерева вычислительное устройство может определить интегральную оценку на основе априорных и вычисляемых оценок. Дерево с наилучшей оценкой может быть выбрано для построения наилучшей синтаксической структуры 746 исходного предложения 212.

[000138] В ходе построения синтаксической структуры 746 на основе выбранного синтаксического дерева вычислительное устройство может установить одну или несколько недревесных связей (например, путем создания дополнительной связи среди, как минимум, двух узлов графа). Если этот процесс заканчивается неудачей, вычислительное устройство может выбрать синтаксическое дерево с условно оптимальной оценкой, наиболее близкой к оптимальной, и производится попытка установить одну или несколько недревесных связей в дереве. Наконец, в результате точного синтаксического анализа создается синтаксическая структура 746, которая представляет собой лучшую синтаксическую структуру, соответствующую исходному предложению 212. Фактически в результате отбора лучшей синтаксической структуры 746 определяются лучшие лексические значения 240 для элементов исходного предложения 212.

[000139] В блоке 217 вычислительное устройство может обрабатывать синтаксические деревья для формирования семантической структуры 218, соответствующей предложению 212. Семантическая структура 218 может отражать передаваемую исходным предложением семантику в независимых от языка терминах. Семантическая структура 218 может быть представлена в виде ациклического графа (например, дерево, возможно, дополненное одной или более недревесной связью (дугой графа). Слова исходного предложения представлены узлами с соответствующими независимыми от языка семантическими классами семантической иерархии 510. Дуги графа представляют глубинные (семантические) отношения между элементами предложения. Переход к семантической структуре 218 может осуществляться с помощью правил анализа 960 и

предполагает соотнесение одного или более атрибутов (отражающих лексические, синтаксические и (или) семантические свойства слов исходного предложения 212) с каждым семантическим классом.

[000140] На Фиг. 14 приводится пример синтаксической структуры предложения, сгенерированной из графа обобщенных составляющих, показанного на Фиг. 13. Узел 901 соответствует лексическому элементу "life" (жизнь) 906. Применяя способ описанного в настоящем документе синтактико-семантического анализа, вычислительное устройство может установить, что лексический элемент "life" (жизнь) 906 представляет одну из форм лексического значения, соотнесенного с семантическим классом "LIVE" (ЖИТЬ) 904 и заполняет поверхностную позицию $Adjunct_Locative 905) в родительской составляющей, представленной управляющим узлом Verb:succeed:succeed:TO_SUCCEED (907).

[000141] На Фиг. 15 приводится семантическая структура, соответствующая синтаксической структуре на Фиг. 15. В отношении вышеупомянутого лексического элемента "life" (жизнь) (906) на Фиг. 13 семантическая структура включает лексический класс 1010 и семантический класс 1030, соответствующие представленным на Фиг. 15, однако вместо поверхностной позиции (905) семантическая структура включает глубинную позицию "Sphere" (сфера_деятельности) 1020.

[000142] Как отмечено выше в настоящем документе, в качестве «онтологии» может выступать модель, которая представляет собой объекты, относящиеся к определенной области знаний (предметной области), и отношения между данными объектами. Таким образом, онтология отличается от семантической иерархии, несмотря на то что она может быть соотнесена с элементами семантической иерархии через определенные отношения (также называемые «якоря»). Онтология может включать определения некого множества классов, где каждый класс соответствует концепту предметной области. Каждое определение класса может включать определения одного или более отнесенных к данному классу объектов. Согласно общепринятой терминологии класс онтологии может также называться «концепт», а принадлежащий классу объект может означать экземпляр данного концепта.

[000143] В соответствии с одним или несколькими аспектами настоящего изобретения вычислительное устройство, в котором реализованы описанные в настоящем описании способы, может индексировать один или несколько параметров, полученных в результате семантико-синтаксического анализа. Таким образом, способы настоящего изобретения позволяют рассматривать не только множество слов в составе исходного текстового корпуса, но и множество лексических значений этих слов, сохраняя и индексируя всю синтаксическую и семантическую информацию, полученную в ходе синтаксического и семантического анализа каждого предложения исходного текстового корпуса. Такая информация может дополнительно включать данные, полученные в ходе промежуточных этапов анализа, а также результаты лексического выбора, в том числе результаты, полученные в ходе разрешения неоднозначностей, вызванных омонимией и(или) совпадающими грамматическими формами, соответствующими различным лексико-морфологическим значениям некоторых слов исходного языка.

[000144] Для каждой семантической структуры можно создать один или несколько индексов. Индекс можно представить в виде структуры данных в памяти, например, в виде таблицы, состоящей из нескольких записей. Каждая запись может представлять собой установление соответствия между определенным элементом семантической структуры (например, одно слово или несколько слов, синтаксическое отношение, морфологическое, синтаксическое или семантическое свойство или синтаксическая или

семантическая структура) и одним или несколькими идентификаторами (или адресами) случаев употребления данного элемента семантической структуры в исходном тексте.

[000145] В некоторых вариантах осуществления индекс может включать одно или несколько значений морфологических, синтаксических, лексических и(или) семантических параметров. Эти значения могут создаваться в процессе двухэтапного семантического анализа (более подробное описание см. в настоящем документе). Индекс можно использовать для выполнения различных задач обработки естественного языка, в том числе для выполнения семантического поиска.

[000146] Вычислительное устройство, реализующее данный способ, может извлекать широкий спектр лексических, грамматических, синтаксических, прагматических и(или) семантических характеристик в ходе проведения синтактико-семантического анализа и создания семантических структур. В иллюстративном примере система может извлекать и сохранять определенную лексическую информацию, данные о принадлежности определенных лексических единиц семантическим классам, информацию о грамматических формах и линейном порядке, информацию об использовании определенных форм, аспектов, тональности {например, положительной или отрицательной), глубинных позиций, недревесных связей, семантем и т.д.

[000147] Вычислительное устройство, в котором реализованы описанные здесь способы, может производить анализ, используя один или несколько описанных в этом документе способов анализа текста, и индексировать любой один или несколько параметров описаний языка, включая лексические значения, семантические классы, граммемы, семантемы и т.д. Индексацию семантического класса можно использовать в различных задачах обработки естественного языка, включая семантический поиск, классификацию, кластеризацию, фильтрацию текста и т.д.. Индексация лексических значений (вместо индексации слов) позволяет искать не только слова и формы слов, но и лексические значения, т.е. слова, имеющие определенные лексические значения. Вычислительное устройство, реализующее способы настоящего изобретения, также может хранить и индексировать синтаксические и семантические структуры, созданные одним или несколькими описанными в настоящем документе способами анализа текста, для использования данных структур и(или) индексов при проведении семантического поиска, классификации, кластеризации и фильтрации документов.

[000148] На Фиг. 16 показан иллюстративный пример вычислительной системы 1000, которая может исполнять набор команд, которые вызывают выполнение вычислительной системой любого отдельно взятого или нескольких способов настоящего изобретения. Вычислительная система может быть соединена с другой вычислительной системой по локальной сети, корпоративной сети, сети экстранет или сети Интернет. Вычислительная система может работать в качестве сервера или клиента в сетевой среде «клиент-сервер» либо в качестве однорангового вычислительного устройства в одноранговой (или распределенной) сетевой среде. Вычислительная система может быть представлена персональным компьютером (ПК), планшетным ПК, телевизионной приставкой (STB), карманным ПК (PDA), сотовым телефоном или любой вычислительной системой, способной выполнять набор команд (последовательно или иным образом), определяющих операции, которые должны быть выполнены этой вычислительной системой. Кроме того, несмотря на то, что показана только одна вычислительная система, термин «вычислительная система» также может включать любую совокупность вычислительных систем, которые отдельно или совместно выполняют набор (или более наборов) команд для выполнения одной или более способов, обсуждаемых в настоящем документе.

[000149] Пример вычислительной системы 1000 включает процессор 502, основное запоминающее устройство 504 {например, постоянное запоминающее устройство (ПЗУ) или динамическое оперативное запоминающее устройство (DRAM)) и устройство хранения данных 518, которые взаимодействуют друг с другом по шине 530.

[000150] Процессор 502 может быть представлен одной или более универсальными вычислительными системами, например, микропроцессором, центральным процессором и т.д. В частности, процессор 502 может представлять собой микропроцессор с полным набором команд (CISC), микропроцессор с сокращенным набором команд (RISC), микропроцессор с командными словами сверхбольшой длины (VLIW), процессор, реализующий другой набор команд или процессоры, реализующие комбинацию наборов команд. Процессор 502 также может представлять собой одну или более вычислительных систем специального назначения, например, заказную интегральную микросхему (ASIC), программируемую пользователем вентильную матрицу (FPGA), процессор цифровых сигналов (DSP), сетевой процессор и т.п. Процессор 502 реализован с возможностью выполнения команд 526 для осуществления рассмотренных в настоящем документе операций и функций.

[000151] Вычислительная система 1000 может дополнительно включать устройство сетевого интерфейса 522, устройство визуального отображения 510, устройство ввода символов 512 (например, клавиатуру) и устройство ввода в виде сенсорного экрана 514.

[000152] Устройство хранения данных 518 может содержать машиночитаемый носитель данных 524, в котором хранится один или более наборов команд 526 и в котором реализованы одна или более методов или функций, рассмотренных в настоящем документе. Команды 526 также могут находиться полностью или по меньшей мере частично в основном запоминающем устройстве 504 и/или в процессоре 502 во время выполнения их в вычислительной системе 1000, при этом оперативное запоминающее устройство 504 и процессор 502 также представляют собой машиночитаемый носитель данных. Команды 526 также могут передаваться или приниматься по сети 516 через устройство сетевого интерфейса 522.

[000153] В некоторых вариантах реализации изобретения набор команд 526 может содержать команды способа 100, 300 и/или 400 для извлечения информации из текстов на естественном языке и обучения классификаторов в соответствии с одним или более вариантами реализации настоящего изобретения. Хотя машиночитаемый носитель данных 524 показан в примере на Фиг. 16 в виде одного носителя, термин «машиночитаемый носитель» следует понимать в широком смысле, подразумевающем один или более носителей (например, централизованную или распределенную базу данных и/или соответствующие кэши и серверы), в которых хранится один или более наборов команд. Термин «машиночитаемый носитель данных» также следует понимать как включающий любой носитель, который может хранить, кодировать или переносить набор команд для выполнения машиной и который обеспечивает выполнение машиной любой одной или более методов настоящего изобретения. Поэтому термин «машиночитаемый носитель данных» относится, помимо прочего, к твердотельным запоминающим устройствам, а также к оптическим и магнитным носителям.

[000154] Способы, компоненты и функции, описанные в этом документе, могут быть реализованы с помощью дискретных компонентов оборудования либо они могут быть встроены в функции других компонентов оборудования, например, ASICS (специализированная заказная интегральная схема), FPGA (программируемая логическая интегральная схема), DSP (цифровой сигнальный процессор) или аналогичных устройств.

Кроме того, способы, компоненты и функции могут быть реализованы с помощью модулей встроенного программного обеспечения или функциональных схем аппаратного обеспечения. Способы, компоненты и функции также могут быть реализованы с помощью любой комбинации аппаратного обеспечения и программных компонентов либо исключительно с помощью программного обеспечения.

[000155] В приведенном выше описании изложены многочисленные детали. Однако любому специалисту в этой области техники, ознакомившемуся с этим описанием, должно быть очевидно, что настоящее изобретение может быть осуществлено на практике без этих конкретных деталей. В некоторых случаях хорошо известные структуры и устройства показаны в виде блок-схем без детализации, чтобы не усложнять описание настоящего изобретения.

[000156] Некоторые части описания предпочтительных вариантов реализации изобретения представлены в виде алгоритмов и символического представления операций с битами данных в запоминающем устройстве компьютера. Такие описания и представления алгоритмов представляют собой средства, используемые специалистами в области обработки данных, что обеспечивает наиболее эффективную передачу сущности работы другим специалистам в данной области. В контексте настоящего описания, как это и принято, алгоритмом называется логически непротиворечивая последовательность операций, приводящих к желаемому результату. Операции подразумевают действия, требующие физических манипуляций с физическими величинами. Обычно, хотя и необязательно, эти величины принимают форму электрических или магнитных сигналов, которые можно хранить, передавать, комбинировать, сравнивать и выполнять другие манипуляции. Иногда удобно, прежде всего для обычного использования, описывать эти сигналы в виде битов, значений, элементов, символов, терминов, цифр и т.д.

[000157] Однако следует иметь в виду, что все эти и подобные термины должны быть связаны с соответствующими физическими величинами и что они являются лишь удобными обозначениями, применяемыми к этим величинам. Если явно не указано обратное, принимается, что в последующем описании термины «определение», «вычисление», «расчет», «получение», «установление», «определение», «изменение» и т.п. относятся к действиям и процессам вычислительной системы или аналогичной электронной вычислительной системы, которая использует и преобразует данные, представленные в виде физических {например, электронных) величин в реестрах и запоминающих устройствах вычислительной системы, в другие данные, также представленные в виде физических величин в запоминающих устройствах или реестрах вычислительной системы или иных устройствах хранения, передачи или отображения такой информации.

[000158] Настоящее изобретение также относится к устройству для выполнения операций, описанных в настоящем документе. Такое устройство может быть специально сконструировано для требуемых целей, либо оно может представлять собой универсальный компьютер, который избирательно приводится в действие или дополнительно настраивается с помощью программы, хранящейся в запоминающем устройстве компьютера. Такая компьютерная программа может храниться на машиночитаемом носителе данных, например, помимо прочего, на диске любого типа, включая дискеты, оптические диски, CD-ROM и магнитно-оптические диски, постоянные запоминающие устройства (ПЗУ), оперативные запоминающие устройства (ОЗУ), СППЗУ, ЭППЗУ, магнитные или оптические карты и носители любого типа, подходящие для хранения электронной информации.

[000159] Следует понимать, что приведенное выше описание призвано иллюстрировать, а не ограничивать сущность изобретения. Специалистам в данной области техники после прочтения и уяснения приведенного выше описания станут очевидны и различные другие варианты реализации изобретения. Исходя из этого область применения изобретения должна определяться с учетом прилагаемой формулы изобретения, а также всех областей применения эквивалентных способов, на которые в равной степени распространяется формула изобретения.

(57) Формула изобретения

1. Способ извлечения информации из текстов на естественном языке, включающий:

обучение классификатора извлечения информации для извлечения первого множества информационных объектов из текста на естественном языке, причем определение классификатора извлечения информации включает один или более гиперпараметров;

получение списка извлеченных информационных объектов путем выполнения функции арбитража конфликтов по отношению к множеству конфликтующих информационных объектов, например, первого информационного объекта из набора конфликтующих информационных объектов, принадлежащего к первому множеству информационных объектов и второго информационного объекта из набора конфликтующих информационных объектов, принадлежащего ко второму множеству информационных объектов, извлеченных из текста на естественном языке;

изменение значений гиперпараметров классификатора извлечения информации; и оптимизацию показателя качества извлечения информации для списка извлеченных информационных объектов путем итеративного повторения операций обучения классификатора извлечения информации, выполнения функции арбитража конфликтов и изменения значений гиперпараметров.

2. Способ по п. 1, отличающийся тем, что выполнение функции арбитража конфликтов дополнительно включает:

обучение классификатора арбитража конфликтов, используемого для реализации функции арбитража конфликтов.

3. Способ по п. 1, отличающийся тем, что обучение классификатора извлечения информации дополнительно включает:

определение набора значений множества параметров классификатора извлечения информации для оптимизации показателя качества, причем параметр качества предоставляется в виде гиперпараметра.

4. Способ по п. 1, отличающийся тем, что обучение классификатора извлечения информации производится с использованием обучающей выборки данных, содержащей аннотированный текст на естественном языке, включающий множество текстовых аннотаций, где каждая текстовая аннотация связана с информационным объектом известной категории.

5. Способ по п. 1, отличающийся тем, что функция арбитража конфликтов выполняет как минимум одно действие из следующих: изменение первого информационного объекта, удаление первого информационного объекта или слияние двух или более информационных объектов из набора конфликтующих информационных объектов.

6. Способ по п. 1, дополнительно включающий:

использование классификатора извлечения информации для выполнения операции обработки естественного языка.

7. Способ по п. 1, дополнительно включающий:

выявление набора конфликтующих информационных объектов путем определения

того, что первая текстовая аннотация, связанная с первым информационным объектом, пересекается со второй текстовой аннотацией, связанной со вторым информационным объектом.

8. Способ по п. 1, дополнительно включающий:

выявление набора конфликтующих информационных объектов путем оценки логического условия, включающего первый атрибут первого информационного объекта и второй атрибут второго информационного объекта.

9. Способ по п. 1, дополнительно включающий:

выявление набора конфликтующих информационных объектов путем определения информационного объекта, имеющего ряд атрибутов определенного типа в количестве выше порогового количества атрибутов определенного типа.

10. Способ по п. 1, отличающийся тем, что выполнение функции арбитража конфликтующих объектов дополнительно включает:

применение множества продукционных правил к множеству атрибутов текста на естественном языке.

11. Способ по п. 1, отличающийся тем, что выполнение функции арбитража конфликтующих объектов дополнительно включает:

применение ко множеству атрибутов текста на естественном языке классификатора «арбитража» конфликтов, на основе которого определяется по меньшей мере одно из следующего: сходство первого информационного объекта и второго информационного объекта, представляющих один и тот же объект, степень уверенности первого информационного объекта или степень уверенности второго информационного объекта.

12. Способ по п. 11, дополнительно включающий:

принятие пользовательского ввода, подтверждающего перечень информационных объектов;

добавление к обучающей выборке данных текста на естественном языке в сопровождении метаданных, включающих определения и текстовые аннотации одного или более информационных объектов перечня информационных объектов;

обучение, с использованием обучающей выборки данных, классификатора «арбитража» конфликтов.

13. Способ по п. 1, дополнительно включающий:

принятие пользовательского ввода, подтверждающего перечень информационных объектов;

добавление к обучающей выборке данных текста на естественном языке в сопровождении метаданных, включающих определения и текстовые аннотации одного или более информационных объектов окончательного перечня информационных объектов;

обучение, с использованием обучающей выборки данных, классификатора извлечения информации.

14. Способ по п. 1, дополнительно включающий:

для каждого информационного объекта из списка извлеченных информационных объектов, отображение идентификатора класса информационного объекта в визуальной ассоциации с соответствующим фрагментом текста естественного языка.

15. Способ извлечения информации из текстов на естественном языке, включающий:

извлечение первого множества информационных объектов из текста на естественном языке с использованием первого метода извлечения информации;

извлечение второго множества информационных объектов из текста на естественном языке с использованием второго метода извлечения информации;

изменение значений одного или нескольких гиперпараметров классификатора арбитража конфликтов;

обучение классификатора арбитража конфликтов, который создает список извлеченных информационных объектов посредством выполнения функции арбитража конфликтов в отношении набора конфликтующих информационных объектов, так что первый информационный объект множества конфликтующих информационных объектов принадлежит первому множеству информационных объектов второй информационный объект множества конфликтующих информационных объектов относится ко второму множеству информационных объектов;

оптимизацию показателя качества извлечения информации, отражающую качество списка извлеченных информационных объектов, итеративно повторяя операции извлечения первого множества информационных объектов, извлечение второго множества информационных объектов, изменение значений гиперпараметров и обучение классификатора арбитража конфликтов.

16. Способ по п. 15, в котором классификатор арбитража конфликтов возвращает по меньшей мере одно из: вероятность того, что первый информационный объект и второй информационный объект представляют один и тот же объект, степень уверенности первого информационного объекта или степень уверенность второго информационного объекта.

17. Способ по п. 15, в котором функция арбитража конфликтов выполняет, по меньшей мере, одно из: изменение первого информационного объекта, удаление первого информационного объекта или объединение двух или более информационных объектов множества конфликтующих информационных объектов.

18. Способ по п. 15, в котором извлечение первого множества информационных объектов дополнительно содержит:

обучение классификатора, обеспечивающего степень ассоциации фрагмента текста естественного языка с заранее определенным классом информационных объектов.

19. Способ по п. 15, в котором извлечение первого множества информационных объектов дополнительно содержит:

применение набора продукционных правил к множеству атрибутов текста на естественном языке.

20. Способ по п. 15, дополнительно содержащий:

использование классификатора арбитража конфликтов для выполнения операции обработки естественного языка.

21. Способ по п. 15, дополнительно содержащий:

получение пользовательского ввода, проверяющего список извлеченных информационных объектов;

добавление к набору обучающих данных текста на естественном языке, сопровождаемого метаданными, содержащими определения и текстовые аннотации одного или нескольких информационных объектов из списка информационных объектов; и

использование обучающего набора данных для обучения классификатора арбитража конфликтов.

22. Система извлечения информации из текстов на естественном языке, включающая:

запоминающее устройство;

процессор, связанный с указанным запоминающим устройством, причем этот процессор выполнен с возможностью:

обучать классификатор извлечения информации для извлечения первого множества

информационных объектов из текста на естественном языке, причем определение классификатора извлечения информации включает один или более гиперпараметров;

получать итоговой список извлеченных информационных объектов путем выполнения функции арбитража конфликтов по отношению к множеству конфликтующих информационных объектов, например, первого информационного объекта из набора конфликтующих информационных объектов, принадлежащего к первому множеству информационных объектов, и второго информационного объекта из набора конфликтующих информационных объектов, принадлежащего ко второму множеству информационных объектов, извлеченных из текста на естественном языке;

изменять значения гиперпараметров классификатора извлечения информации; и

оптимизировать показатель качества извлечения информации для итогового списка извлеченных информационных объектов путем итеративного повторения операций обучения классификатора извлечения информации, выполнения функции арбитража конфликтов и изменения значений гиперпараметров.

23. Система по п. 22, отличающаяся тем, что процессор выполнен с возможностью: использовать классификатор извлечения информации для выполнения операций обработки естественного языка.

24. Постоянный машиночитаемый носитель данных, содержащий исполняемые команды для извлечения информации из текстов на естественном языке, которые при их исполнении побуждают вычислительную систему:

обучать классификатор извлечения информации для извлечения первого множества информационных объектов из текста на естественном языке, причем определение классификатора извлечения информации включает один или более гиперпараметров;

получать список извлеченных информационных объектов путем выполнения функции арбитража конфликтов по отношению к множеству конфликтующих информационных объектов, например, первого информационного объекта из набора конфликтующих информационных объектов, принадлежащего к первому множеству информационных объектов и второго информационного объекта из набора конфликтующих информационных объектов, принадлежащего ко второму множеству информационных объектов, извлеченных из текста на естественном языке;

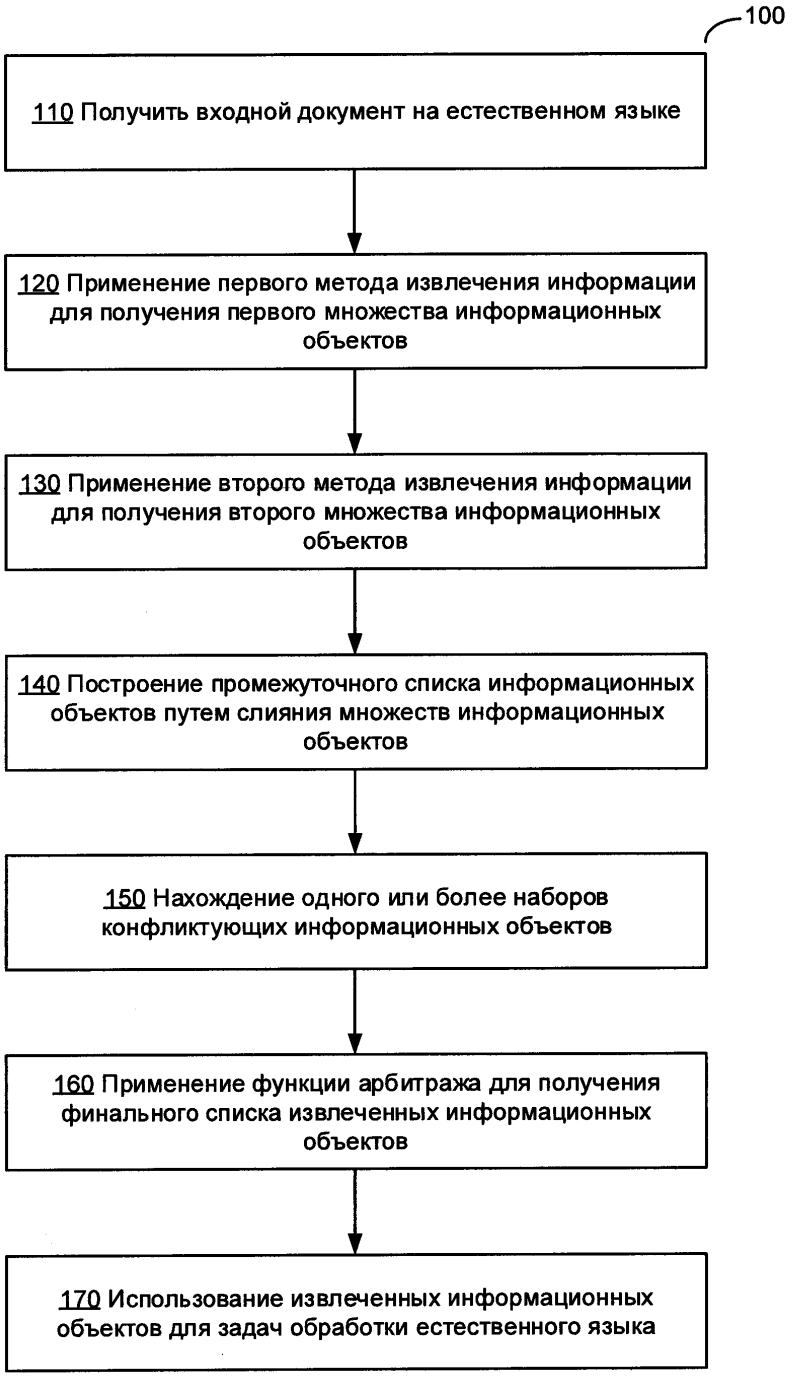
изменять значения гиперпараметров классификатора извлечения информации; и

оптимизировать показатель качества извлечения информации для списка извлеченных информационных объектов путем итеративного повторения операций обучения классификатора извлечения информации, выполнения функции арбитража конфликтов и изменения значений гиперпараметров.

25. Постоянный машиночитаемый носитель данных по п. 24, дополнительно включающий исполняемые команды, которые при выполнении заставляют вычислительную систему:

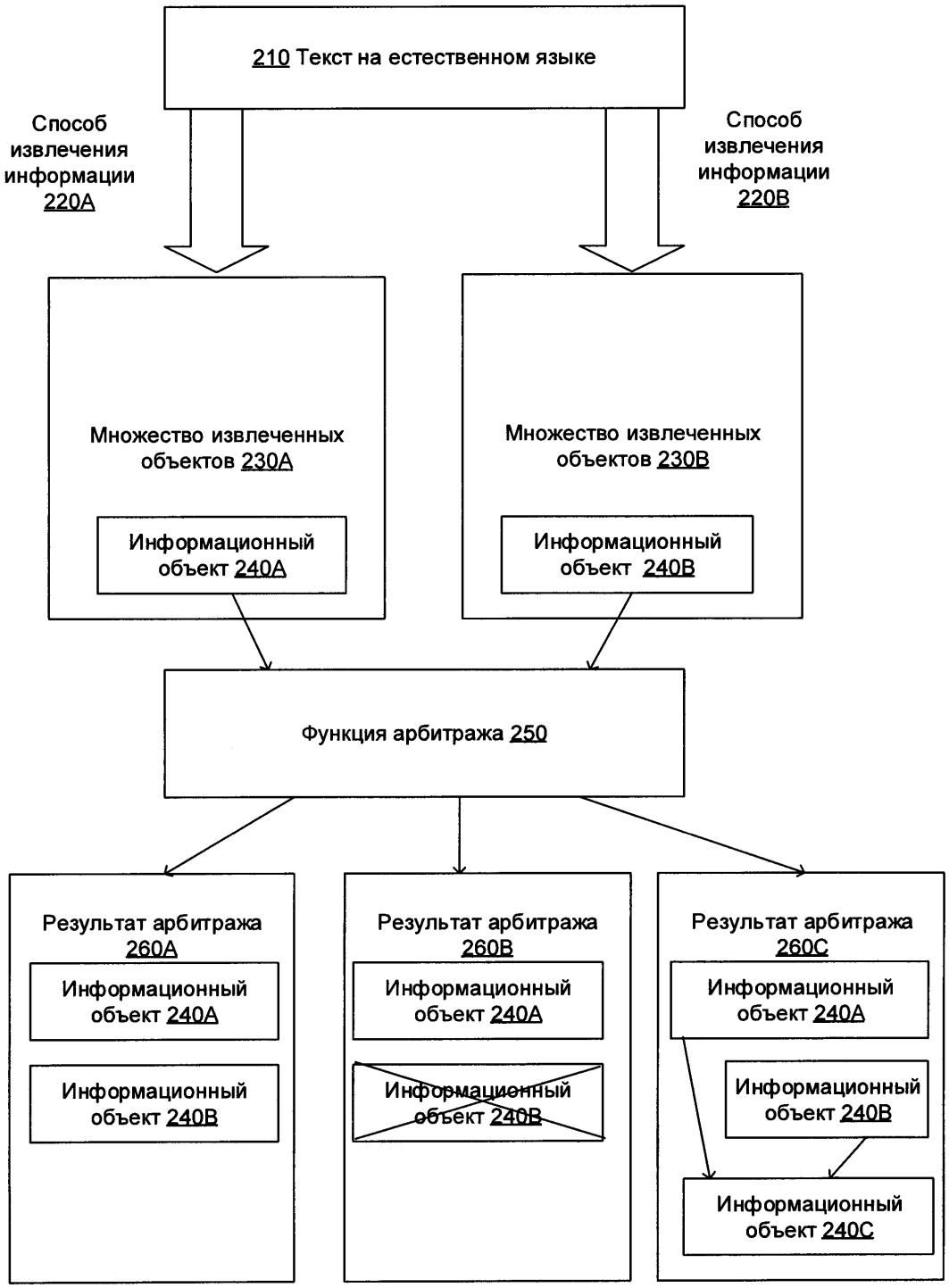
использовать классификатор извлечения информации для выполнения операций обработки естественного языка.

1


100
110 Получить входной документ на естественном языке
120 Применение первого метода извлечения информации для получения первого множества информационных объектов
130 Применение второго метода извлечения информации для получения второго множества информационных объектов
140 Построение промежуточного списка информационных объектов путем слияния множеств информационных объектов
150 Нахождение одного или более наборов конфликтующих информационных объектов
160 Применение функции арбитража для получения финального списка извлеченных информационных объектов
170 Использование извлеченных информационных объектов для задач обработки естественного языка

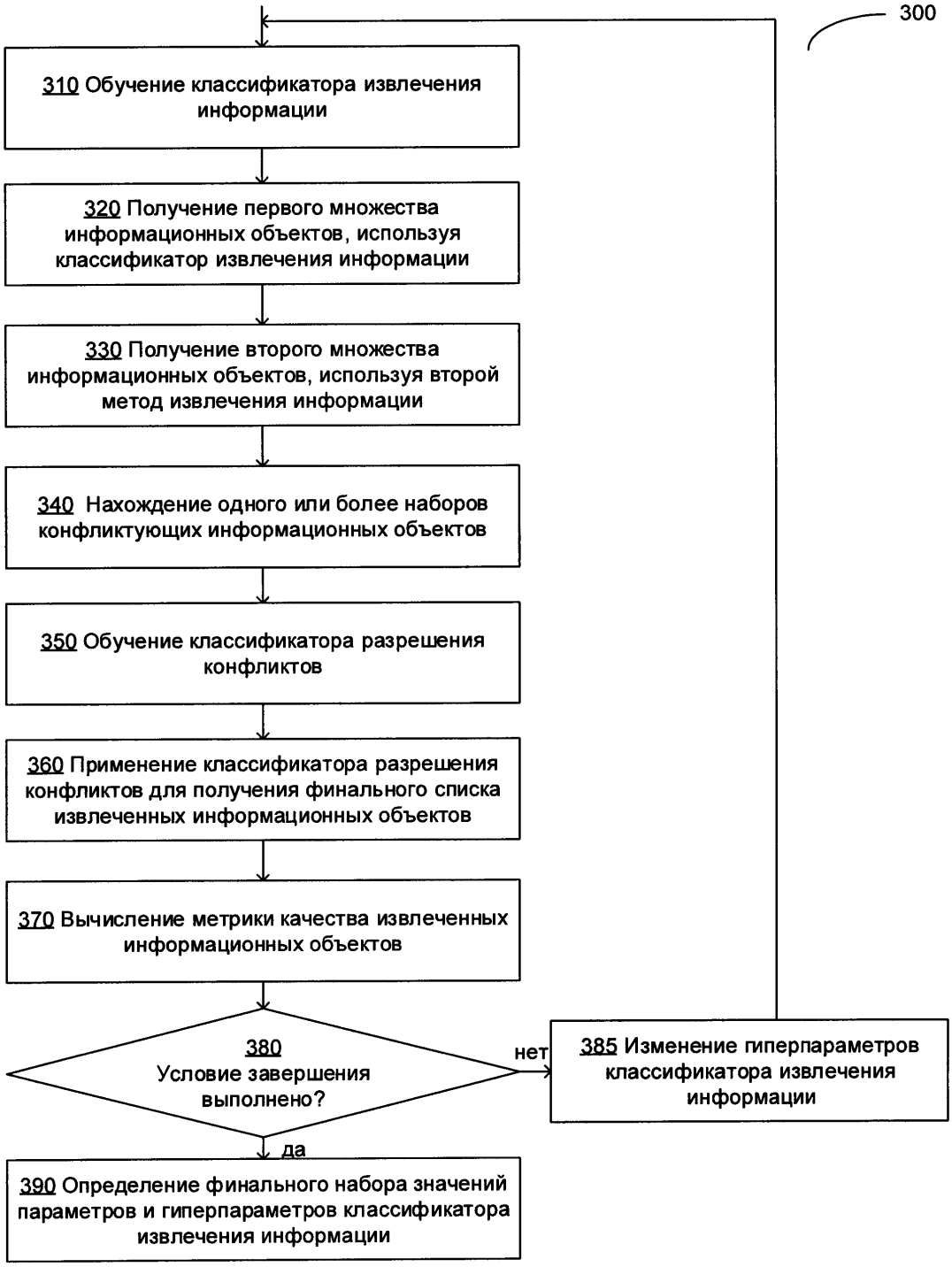
Фиг. 1

2


210 Текст на естественном языке
Способ извлечения информации 220A
Способ извлечения информации 220B
Множество извлеченных объектов 230A
Информационный объект 240A
Множество извлеченных объектов 230B
Информационный объект 240B
Функция арбитража 250
Результат арбитража 260A
Информационный объект 240A
Информационный объект 240B
Результат арбитража 260B
Информационный объект 240A
Информационный объект 240B
Результат арбитража 260C
Информационный объект 240A
Информационный объект 240B
Информационный объект 240C

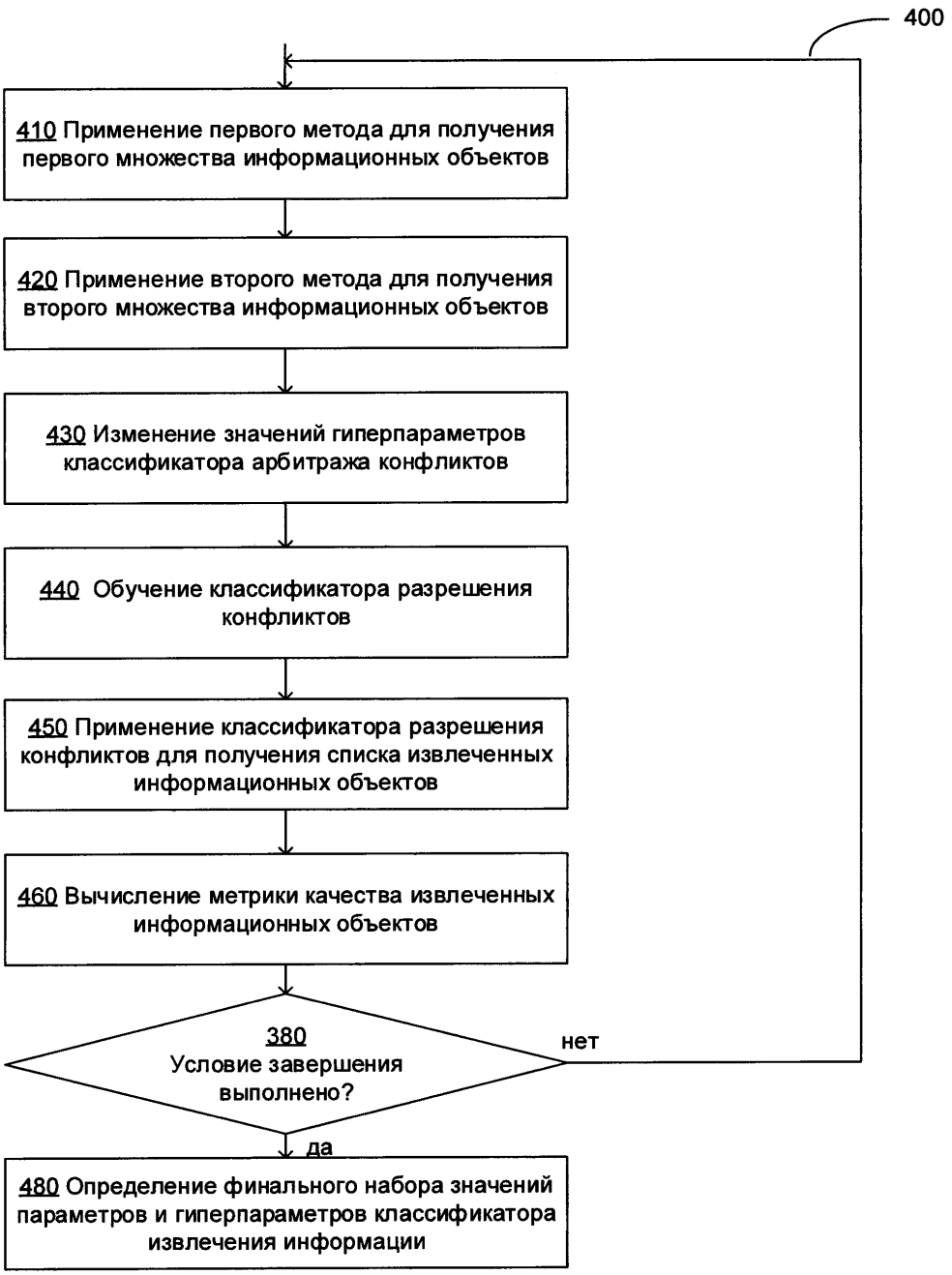
Фиг. 2

3

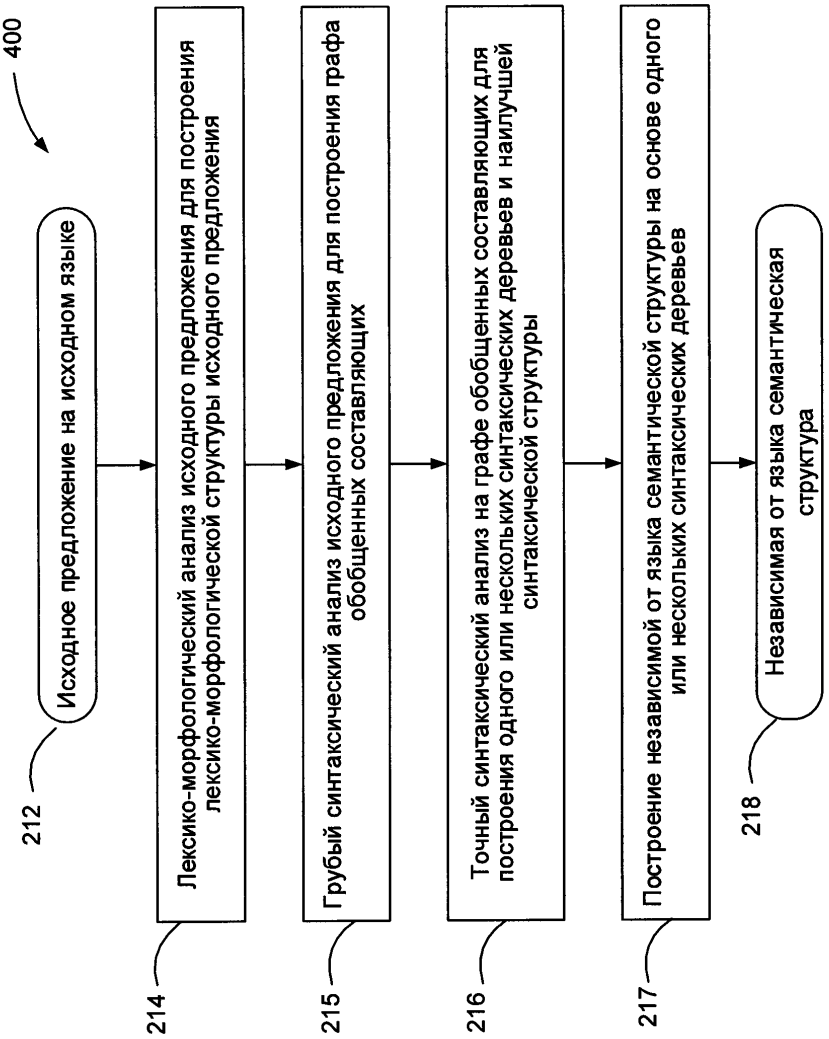

300
310 Обучение классификатора извлечения информации
320 Получение первого множества информационных объектов, используя классификатор извлечения информации
330 Получение второго множества информационных объектов, используя второй метод извлечения информации
340 Нахождение одного или более наборов конфликтующих информационных объектов
350 Обучение классификатора разрешения конфликтов
360 Применение классификатора разрешения конфликтов для получения финального списка извлеченных информационных объектов
370 Вычисление метрики качества извлеченных информационных объектов
380 Условие завершения выполнено?
нет
385 Изменение гиперпараметров классификатора извлечения информации
да
390 Определение финального набора значений параметров и гиперпараметров классификатора извлечения информации

Фиг. 3

4

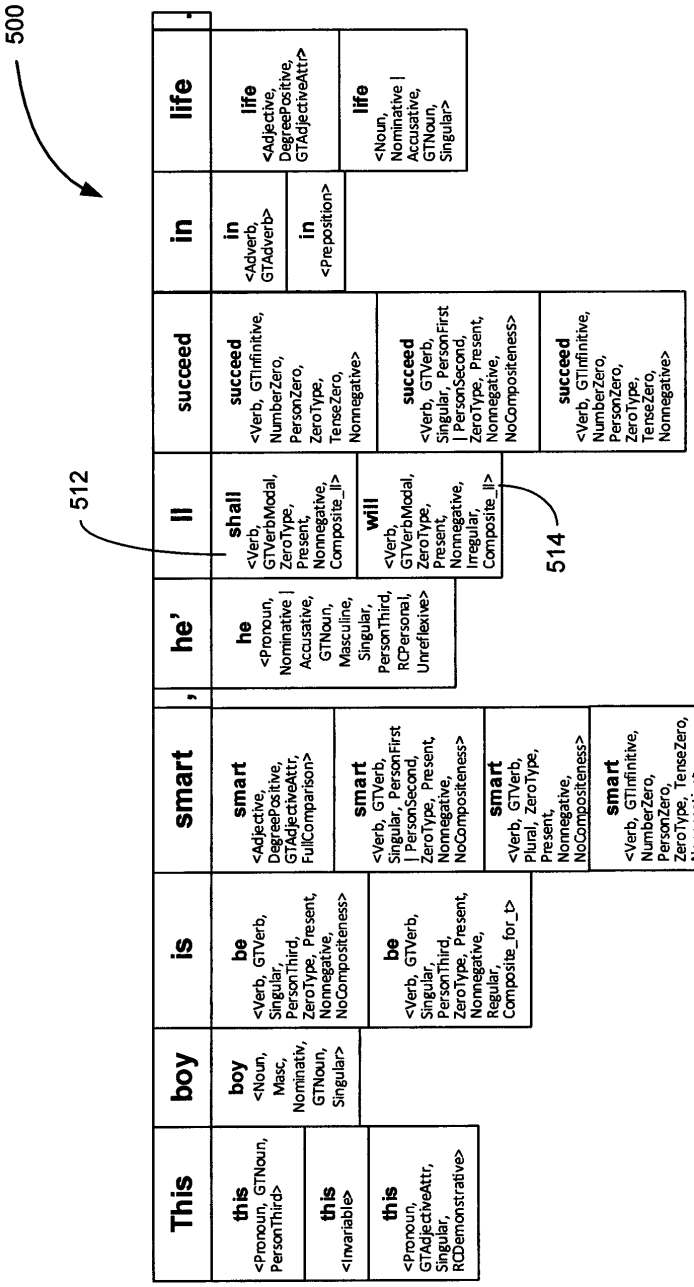

400
410 Применение первого метода для получения первого множества информационных объектов
420 Применение второго метода для получения второго множества информационных объектов
430 Изменение значений гиперпараметров классификатора арбитража конфликтов
440 Обучение классификатора разрешения конфликтов
450 Применение классификатора разрешения конфликтов для получения списка извлеченных информационных объектов
460 Вычисление метрики качества извлеченных информационных объектов
380
Условие завершения выполнено?
нет
да
480 Определение финального набора значений параметров и гиперпараметров классификатора извлечения информации

Фиг. 4

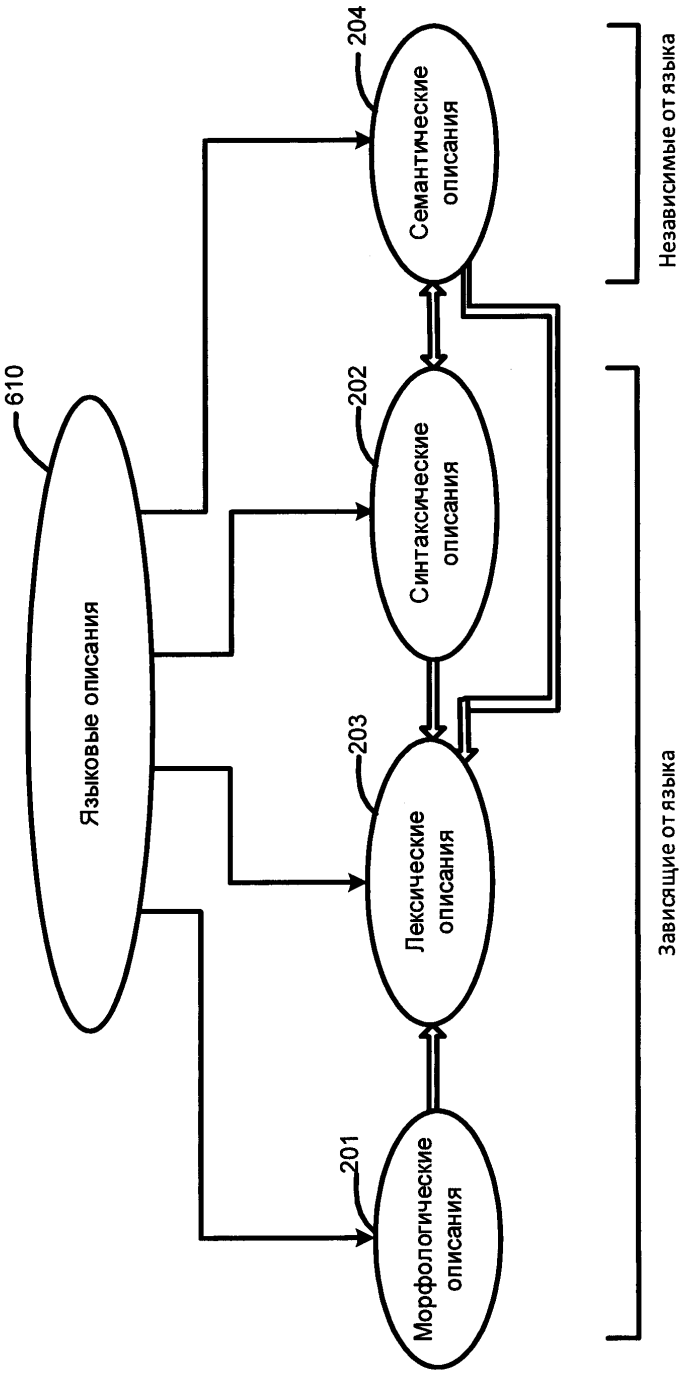

400
Исходное предложение на исходном языке
212
Лексико-морфологический анализ исходного предложения для построения лексико-морфологической структуры исходного предложения
214
Грубый синтаксический анализ исходного предложения для построения графа обобщенных составляющих
215
Точный синтаксический анализ на графе обобщенных составляющих для построения одного или нескольких синтаксических деревьев и наилучшей синтаксической структуры
216
Построение независимой от языка семантической структуры на основе одного или нескольких синтаксических деревьев
217
Независимая от языка семантическая структура
218

Фиг. 5

6

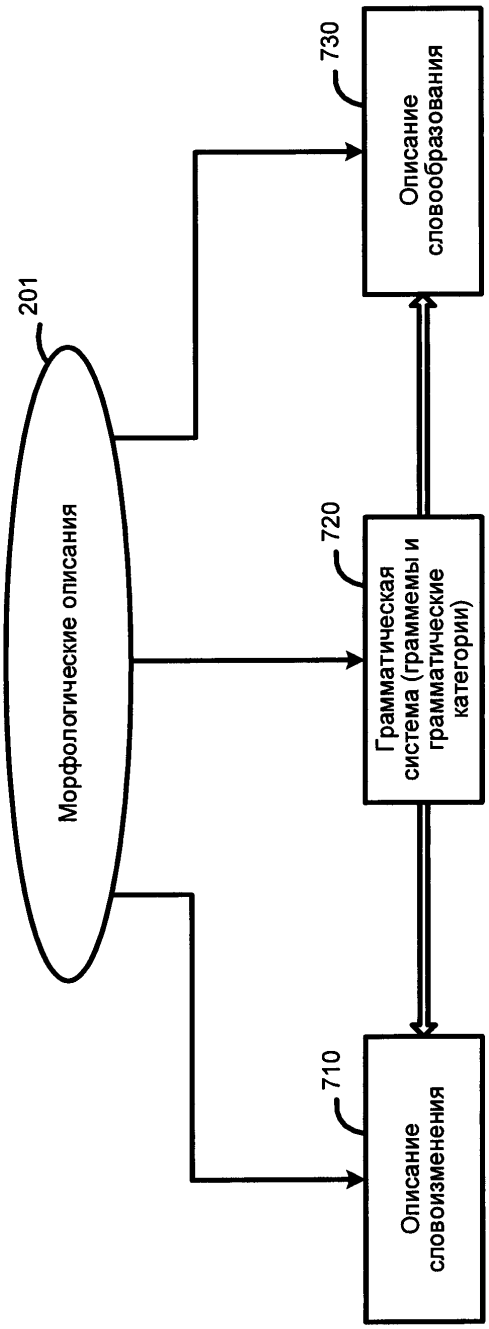

500
512
514
This
boy
is
smart
,
he'
ll
succeed
in
life
.
this
<Pronoun, GTNoun, PersonThird>
this
<Invariable>
this
<Pronoun, GTAdjectiveAttr, Singular, RCDemonstrative>
boy
<Noun, Masc, Nominativ, GTNoun, Singular>
be
<Verb, GTVerb, Singular, PersonThird, ZeroType, Present, Nonnegative, NoCompositeness>
be
<Verb, GTVerb, Singular, PersonThird, ZeroType, Present, Nonnegative, Regular, Composite_for_I>
smart
<Adjective, DegreePositive, GTAdjectiveAttr, FullComparison>
smart
<Verb, GTVerb, Singular, PersonFirst | PersonSecond, ZeroType, Present, Nonnegative, NoCompositeness>
smart
<Verb, GTVerb, Plural, ZeroType, Present, Nonnegative, NoCompositeness>
smart
<Verb, GTInfinitive, NumberZero, PersonZero, ZeroType, TenseZero, Nonnegative>
smart
<Adverb, DegreePositive, GTAdverb, FullComparison>
smart
<Noun, Nominative | Accusative, GTNoun, Singular>
he
<Pronoun, Nominative | Accusative, GTNoun, Masculine, Singular, PersonThird, RCPersonal, Unreflexive>
shall
<Verb, GTVerbModal, ZeroType, Present, Nonnegative, Composite_II>
will
<Verb, GTVerbModal, ZeroType, Present, Nonnegative, Irregular, Composite_II>
succeed
<Verb, GTInfinitive, NumberZero, PersonZero, ZeroType, TenseZero, Nonnegative>
succeed
<Verb, GTVerb, Singular, PersonFirst | PersonSecond, ZeroType, Present, Nonnegative, NoCompositeness>
succeed
<Verb, GTInfinitive, NumberZero, PersonZero, ZeroType, TenseZero, Nonnegative>
in
<Adverb, GTAdverb>
in
<Preposition>
life
<Adjective, DegreePositive, GTAdjectiveAttr>
life
<Noun, Nominative | Accusative, GTNoun, Singular>

Фиг. 6


610
Языковые описания
201
Морфологические описания
203
Лексические описания
202
Синтаксические описания
204
Семантические описания
Независимые от языка
Зависящие от языка

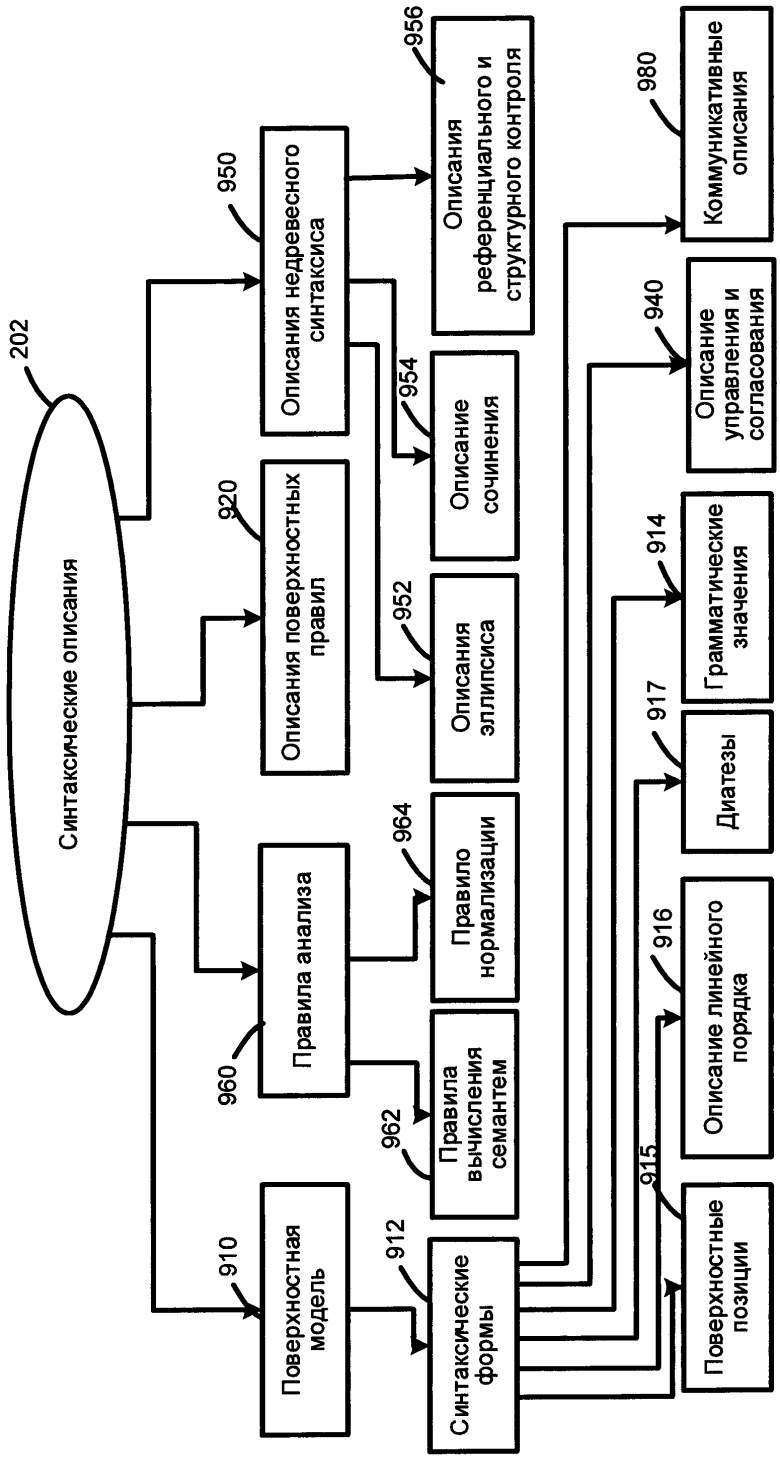
Фиг. 7

8

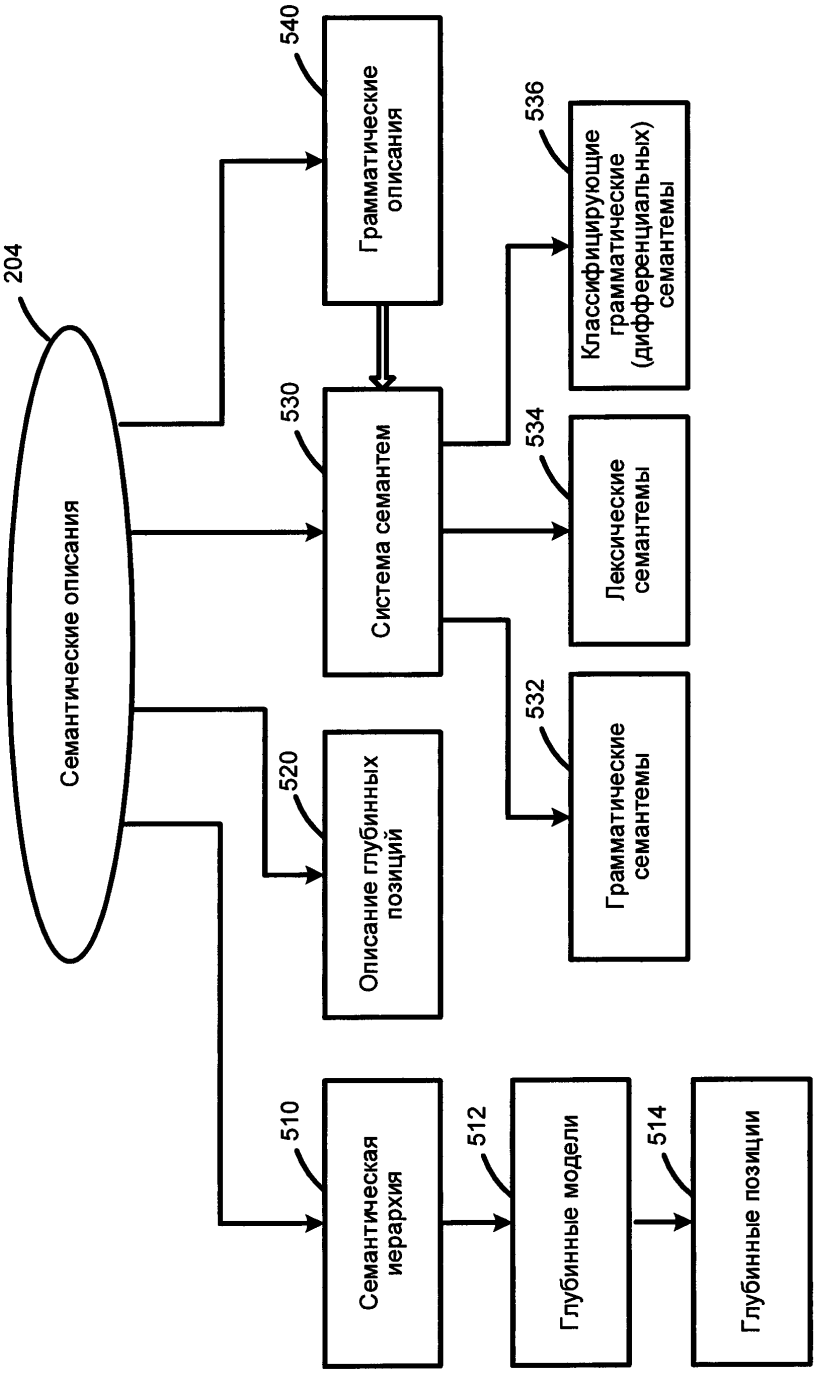

201
Морфологические описания
730
Описание словообразования
720
Грамматическая система (граммемы и грамматические категории)
710
Описание словоизменения

Фиг. 8

9

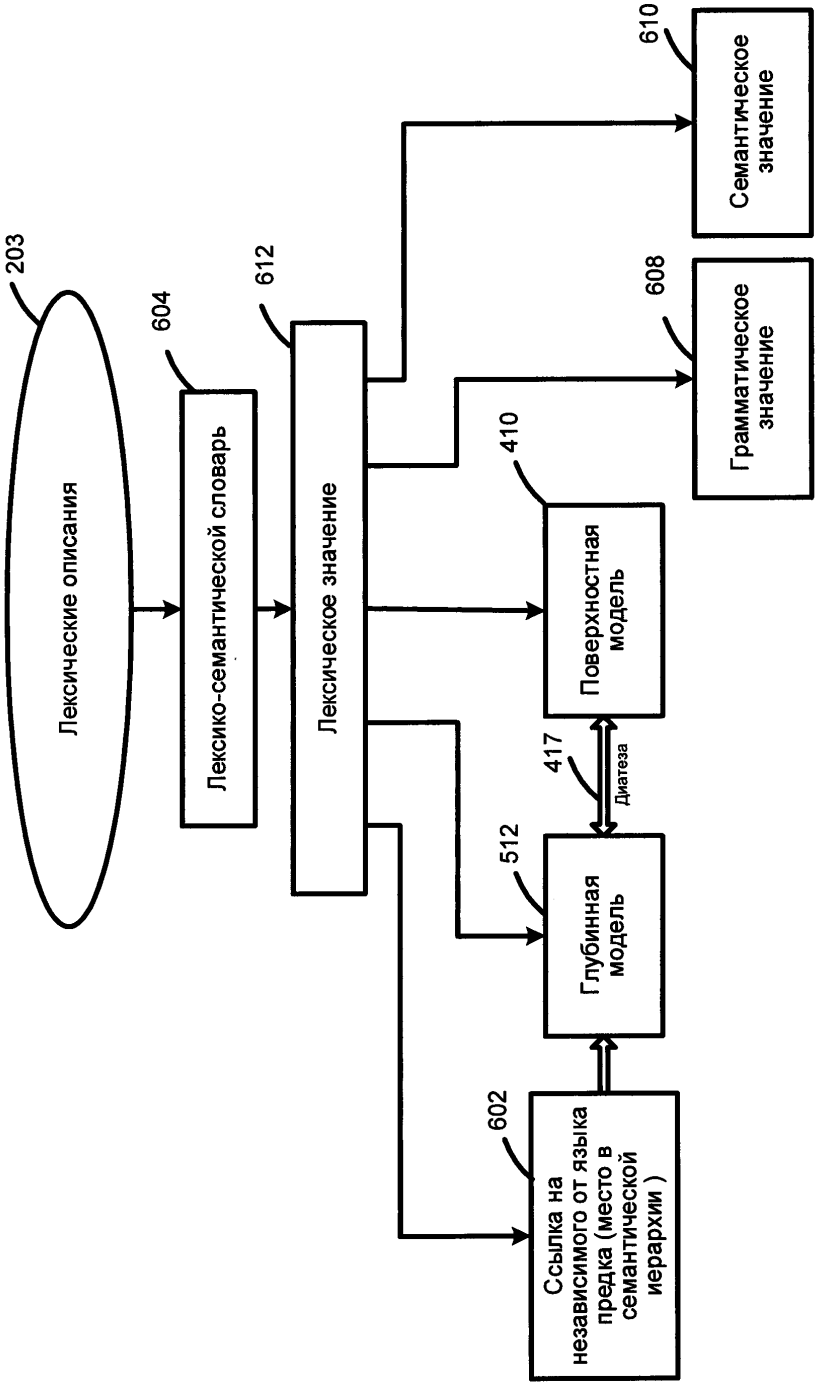

202
Синтаксические описания
910
Поверхностная модель
912
Синтаксические формы
915
Поверхностные позиции
916
Описание линейного порядка
917
Диатезы
914
Грамматические значения
940
Описание управления и согласования
980
Коммуникативные описания
960
Правила анализа
962
Правила вычисления семантем
964
Правило нормализации
920
Описания поверхностных правил
952
Описания эллипсиса
950
Описания недревесного синтаксиса
954
Описание сочинения
956
Описания референциального и структурного контроля

Фиг. 9

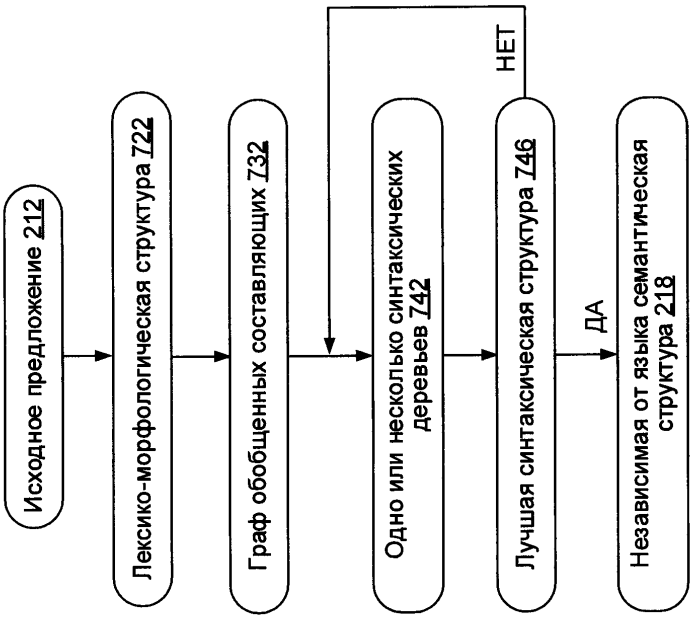

204
Семантические описания
540
Грамматические описания
530
Система семантем
536
Классифицирующие грамматические (дифференциальных) семантемы
534
Лексические семантемы
532
Грамматические семантемы
520
Описание глубинных позиций
510
Семантическая иерархия
512
Глубинные модели
514
Глубинные позиции

Фиг. 10

11

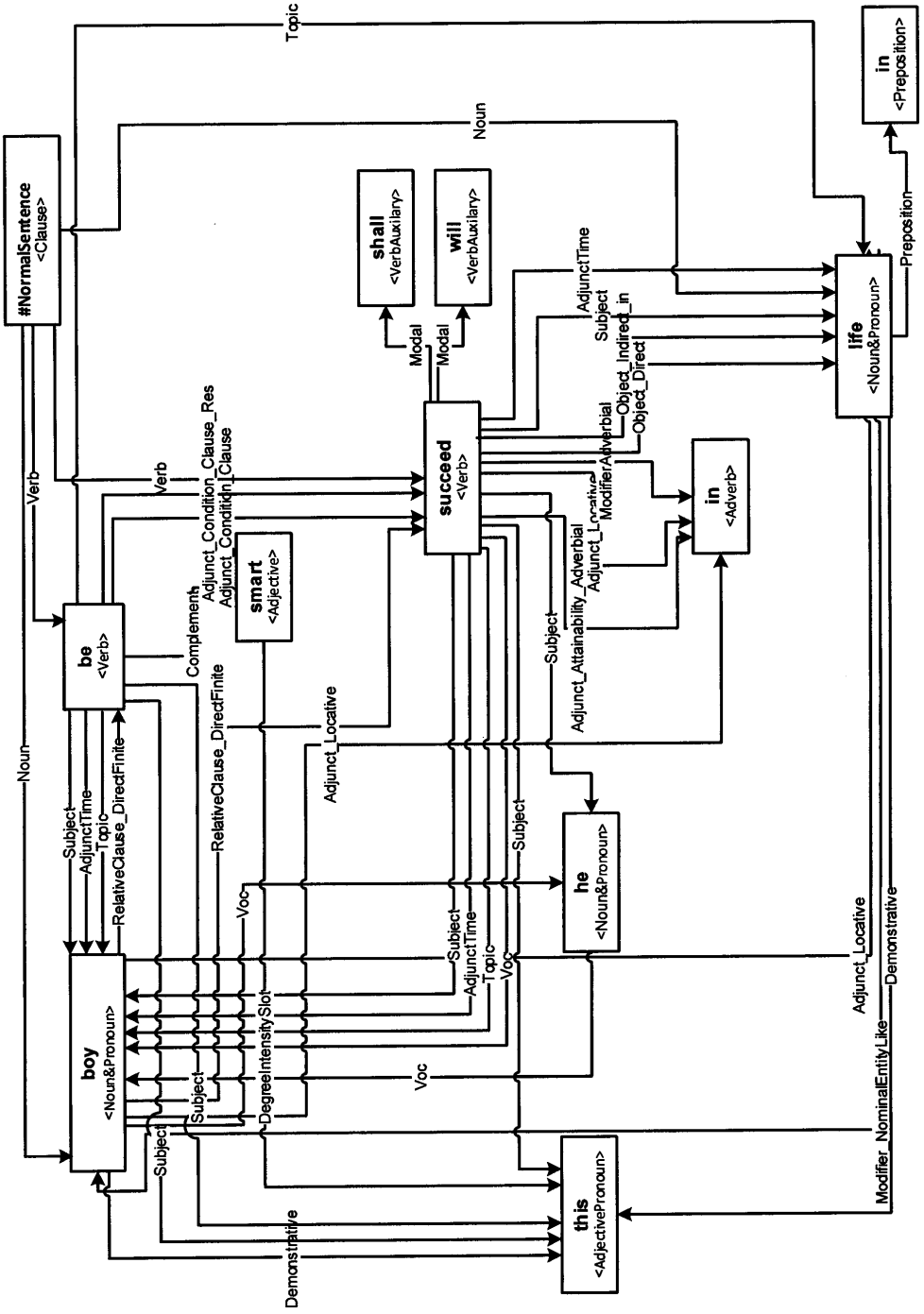

203
Лексические описания
604
Лексико-семантической словарь
612
Лексическое значение
610
Семантическое значение
608
Грамматическое значение
410
Поверхностная модель
417
Диатеза
512
Глубинная модель
602
Ссылка на независимого от языка предка (место в семантической иерархии)

Фиг. 11


Исходное предложение 212
Лексико-морфологическая структура 722
Граф обобщенных составляющих 732
Одно или несколько синтаксических деревьев 742
Лучшая синтаксическая структура 746
НЕТ
ДА
Независимая от языка семантическая структура 218

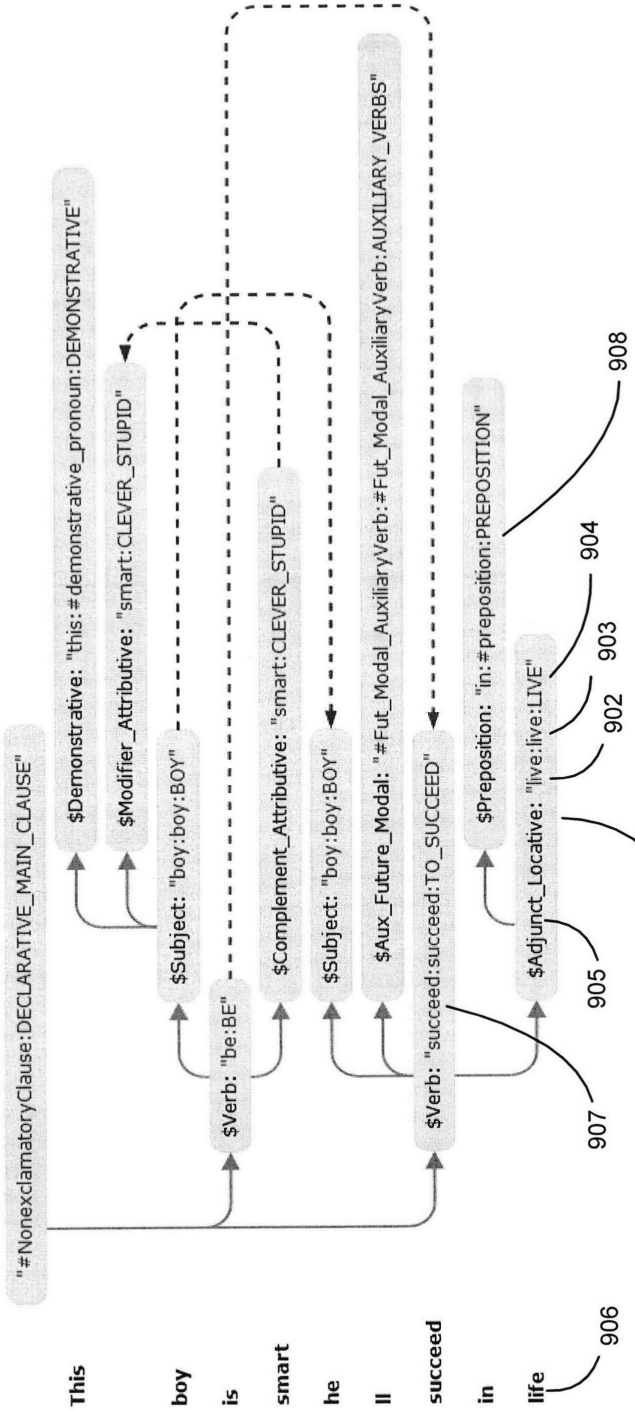
Фиг. 12

13


#NormalSentence
<Clause>
Topic
Noun
Verb
Subject
AdjunctTime
Topic
RelativeClause_DirectFinite
Complement
Adjunct_Condition_Clause_Res
Adjunct_Condition_Clause
RelativeClause_DirectFinite
Adjunct_Locative
Voc
Subject
AdjunctTime
Topic
Voc
Subject
DegreeIntensitySlot
Subject
Voc
Subject
Demonstrative
Modal
Modal
AdjunctTime
Subject
Object_Indirect_in
Object_Direct
ModifierAdverbial
Adjunct_Locative
Adjunct_Attainability_Adverbial
Subject
Preposition
Adjunct_Locative
Demonstrative
Modifier_NominalEntityLike
shall
<VerbAuxilary>
will
<VerbAuxilary>
succeed
<Verb>
smart
<Adjective>
be
<Verb>
boy
<Noun&Pronoun>
he
<Noun&Pronoun>
this
<AdjectivePronoun>
in
<Adverb>
life
<Noun&Pronoun>
in
<Preposition>

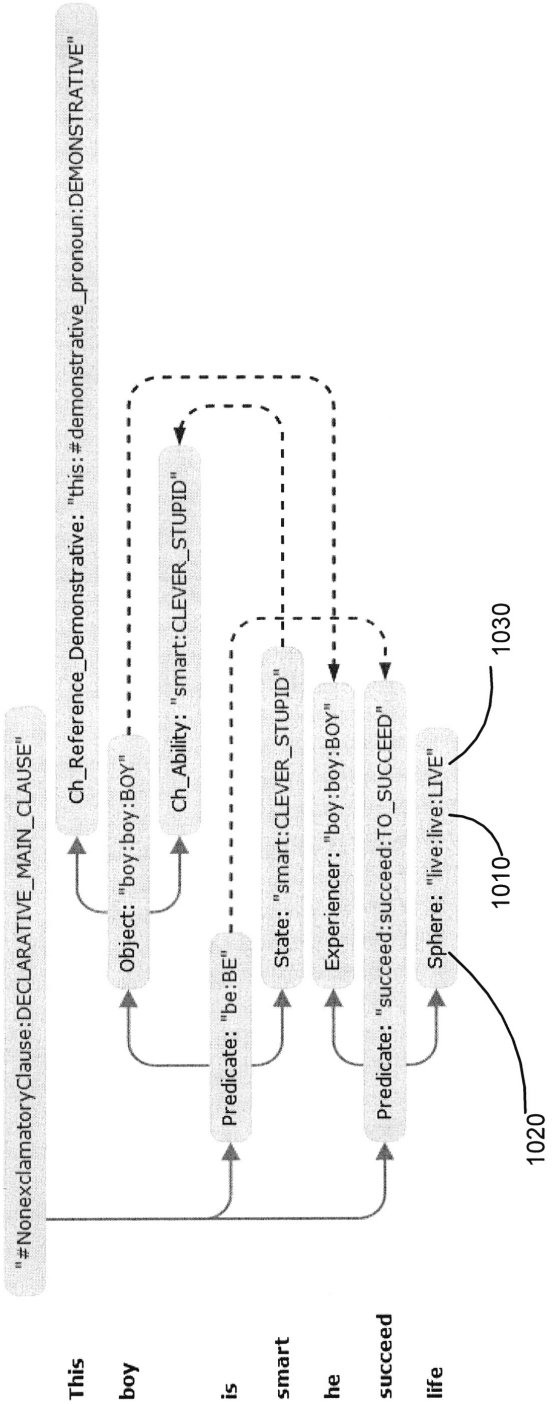
Фиг. 13

14


"#Nonexclamatoryclause:DECLARATIVE_MAIN_CLAUSE"
$Demonstrative: "this:#demonstrative_pronoun:DEMONSTRATIVE"
$Modifier_Attributive: "smart:CLEVER_STUPID"
$Subject: "boy:boy:BOY"
$Verb: "be:BE"
$Complement_Attributive: "smart:CLEVER_STUPID"
$Subject: "boy:boy:BOY"
$Aux_Future_Modal: "#Fut_Modal_AuxiliaryVerb:#Fut_Modal_AuxiliaryVerb:AUXILIARY_VERBS"
$Verb: "succeed:succeed:TO_SUCCEED"
$Preposition: "in:#preposition:PREPOSITION"
$Adjunct_Locative: "live:live:LIVE"
This
boy
is
smart
he
ll
succeed
in
life
901
902
903
904
905
906
907
908

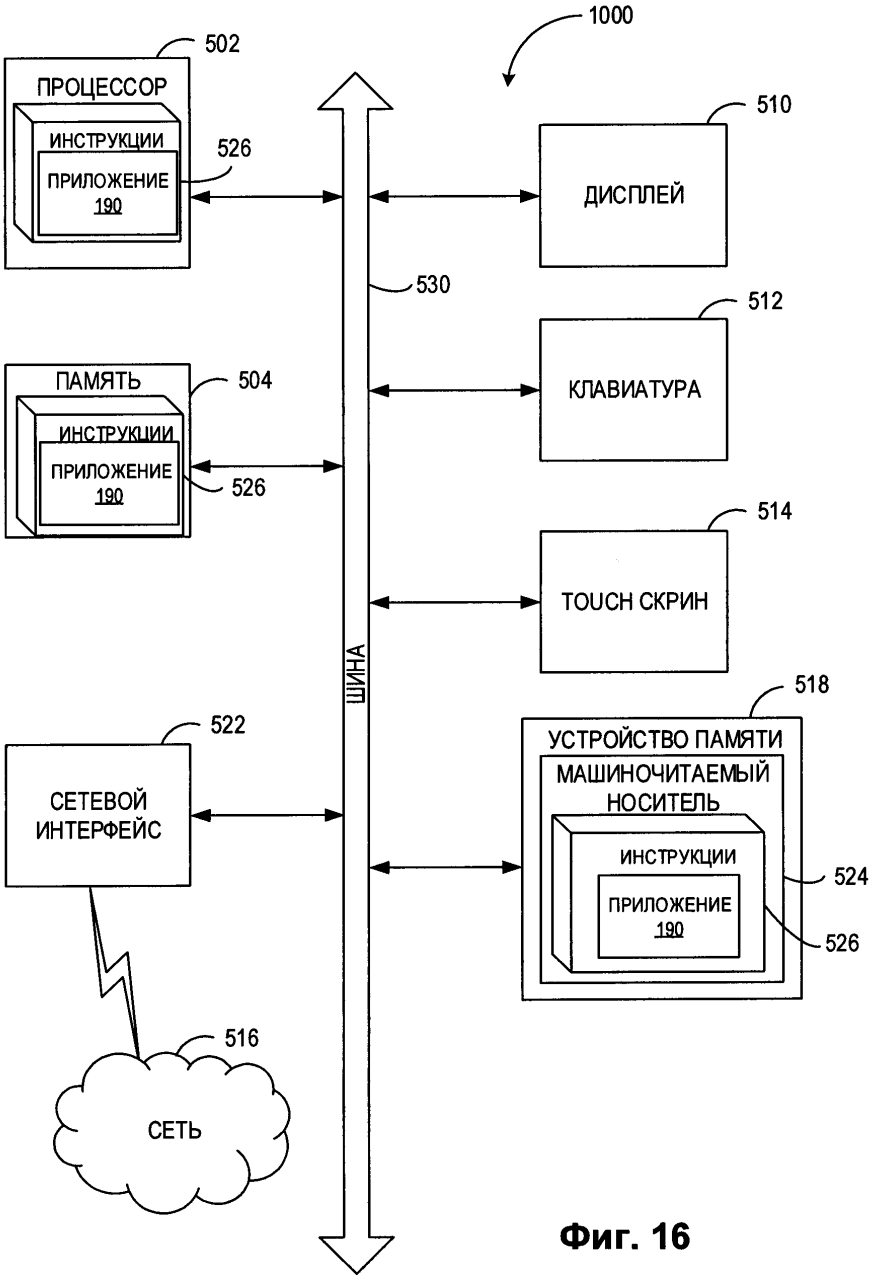
Фиг. 14

15


"#NonexclamatoryClause:DECLARATIVE_MAIN_CLAUSE"
Ch_Reference_Demonstrative: "this:#demonstrative_pronoun:DEMONSTRATIVE"
Object: "boy:boy:BOY"
Ch_Ability: "smart:CLEVER_STUPID"
Predicate: "be:BE"
State: "smart:CLEVER_STUPID"
Experiencer: "boy:boy:BOY"
Predicate: "succeed:succeed:TO_SUCCEED"
Sphere: "live:live:LIVE"
1030
1010
1020
This
boy
is
smart
he
succeed
life

Фиг. 15

16


1000
502
ПРОЦЕССОР
ИНСТРУКЦИИ
ПРИЛОЖЕНИЕ 190
526
510
ДИСПЛЕЙ
530
512
КЛАВИАТУРА
ПАМЯТЬ
504
ИНСТРУКЦИИ
ПРИЛОЖЕНИЕ 190
526
514
TOUCH СКРИН
ШИНА
518
УСТРОЙСТВО ПАМЯТИ
МАШИНОЧИТАЕМЫЙ НОСИТЕЛЬ
ИНСТРУКЦИИ
ПРИЛОЖЕНИЕ 190
524
526
522
СЕТЕВОЙ ИНТЕРФЕЙС
516
СЕТЬ

Фиг. 16