РОССИЙСКАЯ ФЕДЕРАЦИЯ

ФЕДЕРАЛЬНАЯ СЛУЖБА
ПО ИНТЕЛЛЕКТУАЛЬНОЙ СОБСТВЕННОСТИ

(19) **RU** (11) **2 646 386**(13) **C1**

(51) МПК
G06F 17/27 (2006.01)

(12) ОПИСАНИЕ ИЗОБРЕТЕНИЯ К ПАТЕНТУ

(52) СПК
G06F 17/2765 (2006.01); *G06F 17/2785* (2006.01); *G06F 17/28* (2006.01); *G06F 17/30011* (2006.01)

(21)(22) Заявка: **2016147965, 07.12.2016**

(24) Дата начала отсчета срока действия патента:
07.12.2016

Дата регистрации:
02.03.2018

Приоритет(ы):
(22) Дата подачи заявки: **07.12.2016**

(45) Опубликовано: **02.03.2018** Бюл. № **7**

Адрес для переписки:
127273, Москва, а/я 20, ООО "Аби Продакшн", Генеральный директор Терещенко Вадим Владиславович

(72) Автор(ы):
Мацкевич Степан Евгеньевич (RU)

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

(56) Список документов, цитированных в отчете о поиске: **RU 2592396 C1, 20.07.2016. RU 2014101126 A, 20.07.2015. RU 2571373 C2, 20.12.2015. RU 2399959 C2, 20.09.2010. RU 2273879 C2, 10.04.2006. US 7027974 B1, 11.04.2006.**

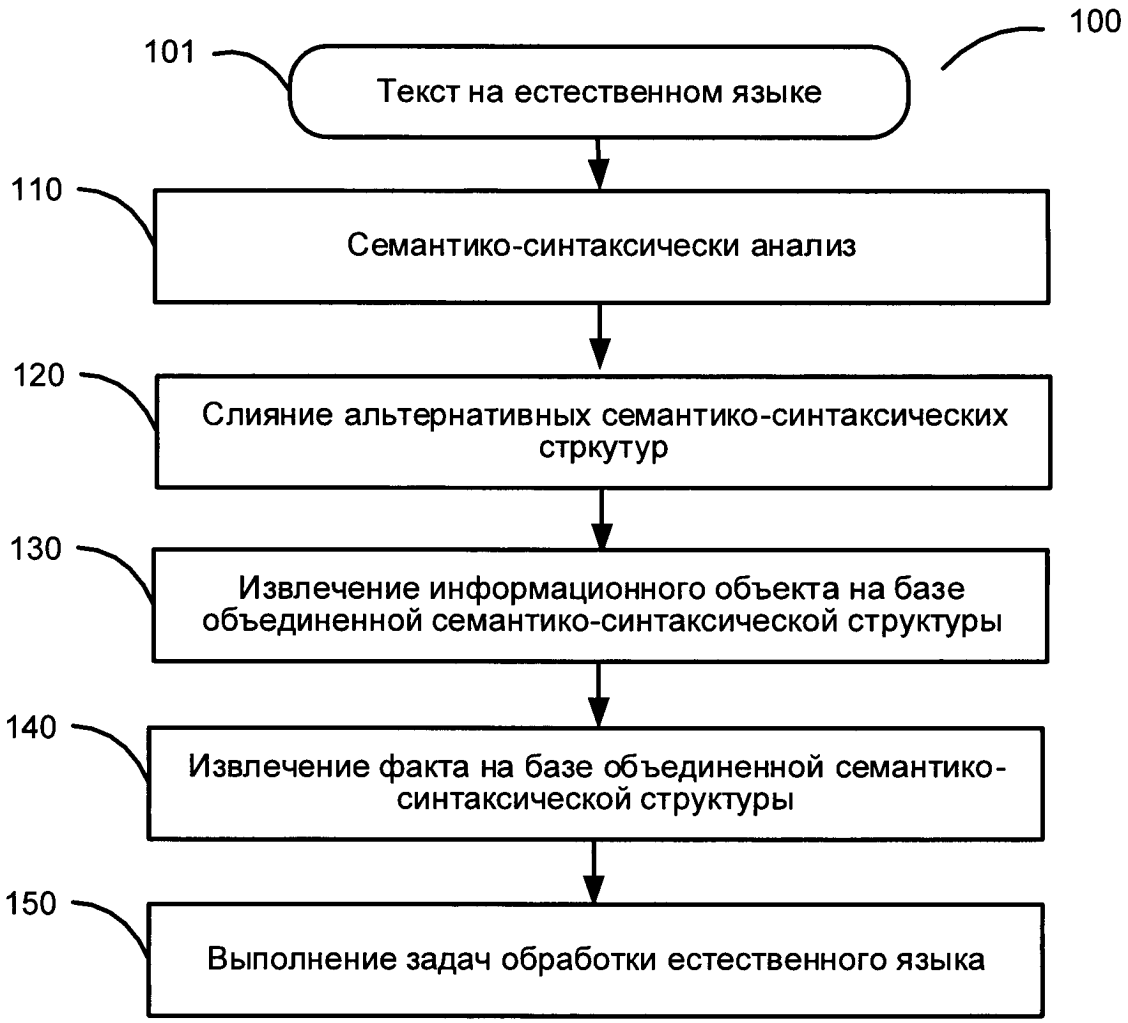
(54) ИЗВЛЕЧЕНИЕ ИНФОРМАЦИИ С ИСПОЛЬЗОВАНИЕМ АЛЬТЕРНАТИВНЫХ ВАРИАНТОВ СЕМАНТИКО-СИНТАКСИЧЕСКОГО РАЗБОРА

(57) Реферат:

Изобретение относится к обработке текстов на естественном языке. Техническим результатом является повышение объема извлечения информации с учетом возможной неоднозначности предложений естественного языка и альтернативных вариантов семантико-синтаксического разбора. В способе извлечения информации из текстов на естественном языке выполняют семантико-синтаксический анализ части текста на естественном языке с целью получения множества семантико-синтаксических структур, включающего первую и вторую альтернативные семантико-синтаксические структуры. Объединяют множество структур с целью получения объединенной семантико-синтаксической структуры. Исключают дублирующие семантико-синтаксические подструктуры из объединенной структуры. Выявляют в пределах указанной части текста информационные объекты путем интерпретации объединенной структуры с целью установления ассоциативной связи токенов, образованных указанной частью текста, с некоторой категорией информационных объектов. При этом интерпретация объединенной структуры производится с учетом значения метрики качества, ассоциированной с частью первой альтернативной структуры. 3 н. и 13 з.п. ф-лы, 13 ил.

RU 2646386 C1

RU 2646386 C1



Фиг. 1

RU 2646386 C1

RU 2646386 C1

RUSSIAN FEDERATION

FEDERAL SERVICE
FOR INTELLECTUAL PROPERTY

(19) **RU** (11) **2 646 386**(13) **C1**

(51) Int. Cl.
G06F 17/27 (2006.01)

(12) ABSTRACT OF INVENTION

(52) CPC
G06F 17/2765 (2006.01); *G06F 17/2785* (2006.01); *G06F 17/28* (2006.01); *G06F 17/30011* (2006.01)

(21)(22) Application: **2016147965, 07.12.2016**

(24) Effective date for property rights:
07.12.2016

Registration date:
02.03.2018

Priority:
(22) Date of filing: **07.12.2016**

(45) Date of publication: **02.03.2018** Bull. № **7**

Mail address:
127273, Moskva, a/ya 20, OOO "Abi Prodakshn", Generalnyj direktor Tereshchenko Vadim Vladislavovich

(72) Inventor(s):
Matskevich Stepan Evgenevich (RU)

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

(54) EXTRACTION OF INFORMATION USING ALTERNATIVE VARIANTS OF SEMANTIC-SYNTACTIC ANALYSIS

(57) Abstract:

FIELD: data processing.

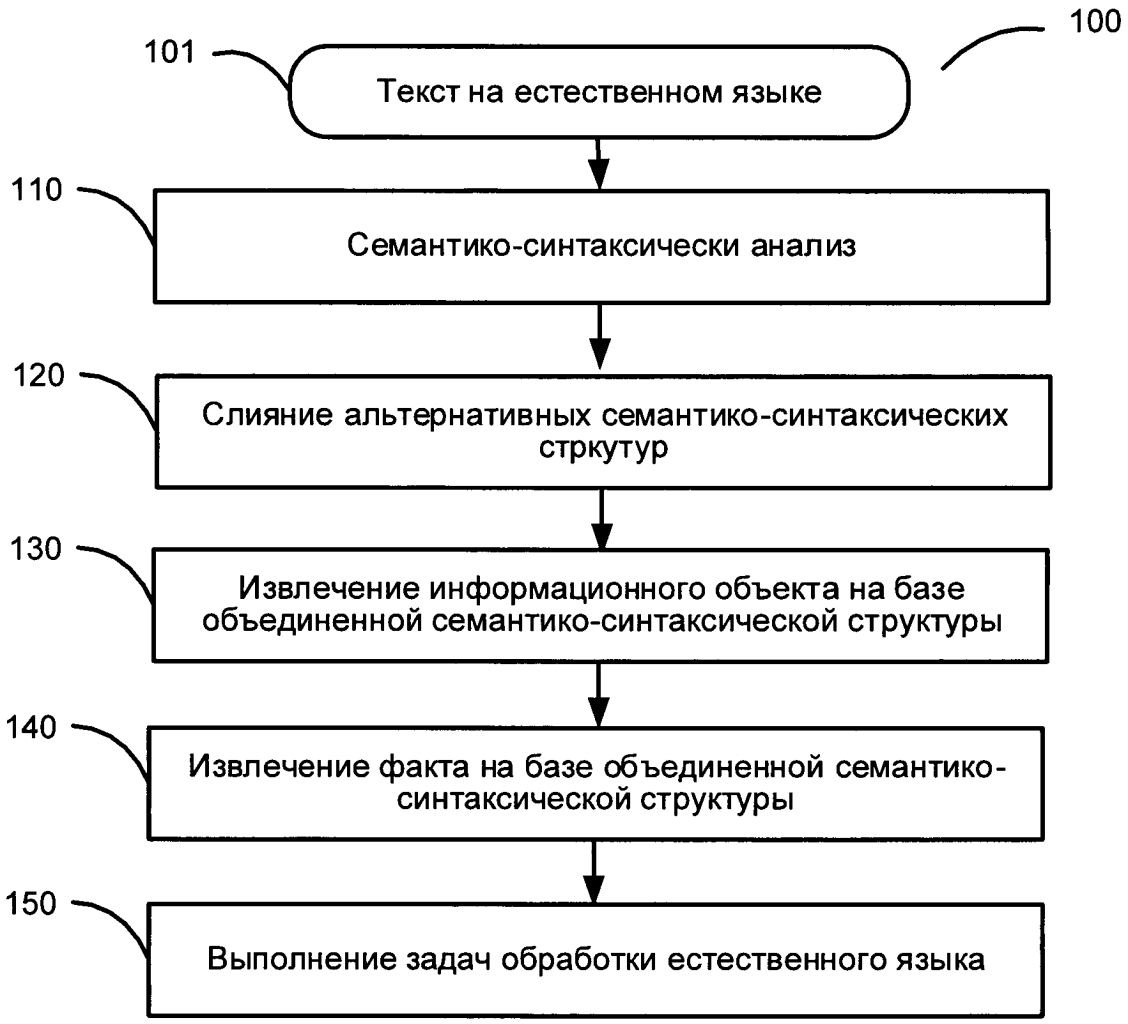
SUBSTANCE: invention relates to the processing of texts in a natural language. In the method of extracting information from texts in a natural language, a semantic-syntactic analysis of a part of the text in a natural language is performed to obtain a multitude of semantic-syntactic structures, including the first and second alternative semantic-syntactic structures. Many structures are combined to obtain a unified semantic-syntactic structure. Duplicating semantic-syntactic substructures are excluded from the combined structure. Information objects are identified within the specified part of the text by interpreting the unified structure in order to establish the associative connection of tokens, formed by this part of the text with a certain category of information objects. In this case, the interpretation of the combined structure is made, taking into account the value of the quality metric associated with the part of the first alternative structure.

EFFECT: technical result is an increase in the volume of information extraction, taking into account the possible ambiguity of sentences in a natural language and alternative variants of semantic-syntactic analysis.

16 cl, 13 dwg

RU 2646386 C1

RU 2646386 C1



Фиг. 1

RU 2646386 C1

RU 2646386 C1

ОБЛАСТЬ ИЗОБРЕТЕНИЯ

[0001] Настоящее изобретение в целом относится к обработке текстов на естественном языке, а в частности - к извлечению информации с учетом альтернативных вариантов семантико-синтаксического разбора.

ПРЕДПОСЫЛКИ СОЗДАНИЯ ИЗОБРЕТЕНИЯ

[0002] Извлечение информации может предусматривать анализ текста на естественном языке с целью выявления информационных объектов, например, именованных сущностей и отношений между выявленными именованными сущностями и другими информационными объектами.

КРАТКОЕ ИЗЛОЖЕНИЕ СУЩНОСТИ ИЗОБРЕТЕНИЯ

[0003] Согласно одному или более вариантам реализации настоящего изобретения предлагается способ извлечения информации с использованием альтернативных вариантов семантико-синтаксического разбора. Пример реализации способа может предусматривать: выполнение вычислительной системой семантико-синтаксического анализа по меньшей мере некоторой части текста на естественном языке с целью получения множества семантико-синтаксических структур, представляющих указанную часть текста, при этом множество семантико-синтаксических структур включает первую альтернативную семантико-синтаксическую структуру и вторую альтернативную семантико-синтаксическую структуру; объединение множества семантико-синтаксических структур с целью получения объединенной семантико-синтаксической структуры; выявление в пределах указанной части текста на естественном языке одного или более информационных объектов путем интерпретации объединенной семантико-синтаксической структуры для установления ассоциативной связи одного или более токенов принадлежащих указанной части текста, с некоторой категорией информационных объектов.

[0004] Согласно одному или нескольким вариантам реализации настоящего изобретения предлагается система извлечения информации с использованием альтернативных вариантов семантико-синтаксического разбора. Пример системы может представлять собой память и процессор, соединенный с памятью, при этом процессор рассчитан на выполнение следующих действий: семантико-синтаксический анализ по меньшей мере некоторой части текста на естественном языке с целью получения множества семантико-синтаксических структур, представляющих часть текста на естественном языке, при этом множество семантико-синтаксических структур включает первую альтернативную семантико-синтаксическую структуру и вторую альтернативную семантико-синтаксическую структуру; объединение множества семантико-синтаксических структур с целью получения объединенной семантико-синтаксической структуры; выявление в пределах указанной части текста на естественном языке одного или более информационных объектов путем интерпретации объединенной семантико-синтаксической структуры с целью установления ассоциативной связи одного или более токенов, образованных указанной частью текста на естественном языке, с некоторой категорией информационных объектов.

[0005] Согласно одному или нескольким вариантам реализации настоящего изобретения предлагается постоянный машиночитаемый носитель данных. Постоянный машиночитаемый носитель данных может предусматривать набор исполняемых команд, которые, при их исполнении на вычислительной системе, обеспечивают выполнение вычислительной системой следующих действий: семантико-синтаксический анализ по меньшей мере некоторой части текста на естественном языке с целью получения множества семантико-синтаксических структур, представляющих часть текста на

естественном языке, при этом множество семантико-синтаксических структур включает первую альтернативную семантико-синтаксическую структуру и вторую альтернативную семантико-синтаксическую структуру; слияние множества семантико-синтаксических структур с целью получения объединенной семантико-синтаксической структуры; выявление в пределах указанной части текста на естественном языке одного или более информационных объектов путем интерпретации объединенной семантико-синтаксической структуры с целью установления ассоциативной связи одного или более токенов, образованных указанной частью текста на естественном языке, с некоторой категорией информационных объектов. Технический результат от внедрения изобретения состоит в возможности более полного извлечения информации с учетом возможной неоднозначности предложений естественного языка, при этом могут учитываться альтернативные варианты семантико-синтаксического разбора.

КРАТКОЕ ОПИСАНИЕ ЧЕРТЕЖЕЙ

[0006] Настоящее изобретение иллюстрируется на примерах, без каких бы то ни было ограничений; его сущность становится понятной при рассмотрении приведенного ниже подробного описания изобретения в сочетании с чертежами, при этом:

[0007] На Фиг. 1 изображена блок-схема примера реализации способа извлечения информации с использованием альтернативных вариантов семантико-синтаксического разбора в соответствии с одним или более вариантами реализации настоящего изобретения.

[0008] На Фиг. 2 приведена блок-схема одного описанного в иллюстративном примере способа выполнения семантико-синтаксического анализа предложения на естественном языке в соответствии с одним или более вариантами реализации настоящего изобретения.

[0009] На Фиг. 3 схематически показан пример лексико-морфологической структуры предложения в соответствии с одним или более вариантами реализации настоящего изобретения.

[00010] На Фиг. 4 схематически показаны языковые описания, представляющие модель естественного языка в соответствии с одним или более вариантами реализации настоящего изобретения.

[00011] На Фиг. 5 схематически показаны примеры морфологических описаний в соответствии с одним или более вариантами реализации настоящего изобретения.

[00012] На Фиг. 6 схематически показаны примеры синтаксических описаний в соответствии с одним или более вариантами реализации настоящего изобретения.

[00013] На Фиг. 7 схематически показаны примеры семантических описаний в соответствии с одним или более вариантами реализации настоящего изобретения.

[00014] На Фиг. 8 схематически показаны примеры лексических описаний в соответствии с одним или более вариантами реализации настоящего изобретения.

[00015] На Фиг. 9 схематически показаны примеры структур данных, которые могут использоваться на практике при воплощении одного или более способов в соответствии с одним или более вариантами реализации настоящего изобретения.

[00016] На Фиг. 10 схематически показан пример графа обобщенных составляющих в соответствии с одним или более вариантами реализации настоящего изобретения.

[00017] На Фиг. 11 дан пример синтаксической структуры, соответствующей предложению, приведенному в качестве примера на Фиг. 10.

[00018] На Фиг. 12 изображена одна из семантических структур, соответствующих синтаксической структуре, представленной на Фиг. 11.

[00019] На Фиг. 13 изображена схема описанной в примере вычислительной системы, в которой реализованы способы, изложенные в настоящем описании изобретения.

ОПИСАНИЕ ПРЕДПОЧТИТЕЛЬНЫХ ВАРИАНТОВ РЕАЛИЗАЦИИ

[00020] В настоящем документе описаны способы и системы извлечения информации с учетом альтернативных вариантов семантико-синтаксического анализа. Системы и способы, представленные в настоящем документе, могут найти применение в самых разных приложениях, 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.

[00021] Термин «вычислительная система» в контексте настоящего документа означает устройство обработки данных, оснащенное универсальным процессором, памятью и по меньшей мере одним интерфейсом связи. Примерами вычислительных систем, в которых могут использоваться способы, описанные в настоящем документе, являются, в частности, настольные компьютеры, ноутбуки, планшетные компьютеры и смартфоны.

[00022] Системы и способы, изложенные в настоящем описании изобретения, способствуют более полному извлечению информации с использованием альтернативных вариантов семантико-синтаксического разбора, как подробнее описано ниже в настоящем документе. В некоторых вариантах реализации изобретения способ реализации 100 вычислительной системы может обеспечивать выполнение семантико-синтаксического анализа исходного текста на естественном языке с целью создания множества семантических структур, представляющих собой предложения из текста на естественном языке.

[00023] Некоторые этапы семантико-синтаксического анализа могут порождать множественные альтернативные результаты - как промежуточные, так и финальные. К примеру, ввиду омонимии и (или) совпадения грамматических форм, соответствующих разным лексико-морфологическим значениям определенного слова, для данного слова в ходе лексико-морфологического анализа может быть установлено множество морфологических значений. Кроме того, в ходе грубого синтаксического анализа могут применяться множественные синтаксические модели, связанные с конкретным предложением, что может приводить ко множественным версиям итогового графа обобщенных составляющих. Таким образом, результатом точного синтаксического анализа может быть некоторое множество синтаксических деревьев, соответствующих заданному первоначальному предложению.

[00024] Известны методы, которые могут предусматривать использование функции метрики качества, с учетом совместимости лексических значений слов в исходном предложении, поверхностных отношений, глубинных отношений и т.д., с целью выбора наилучшего из синтаксических деревьев, соответствующих исходному предложению. Вместе с тем системы и способы, изложенные в настоящем описании изобретения, расширяют функциональность универсальных или специализированных вычислительных устройств, повышая полноту извлечения информации за счет рассмотрения множественности синтаксических деревьев. Это достигается благодаря учету множества альтернативных семантико-синтаксических структур, соответствующих фрагментам исходного текста на естественном языке.

[00025] В некоторых вариантах реализации изобретения, с целью повышения эффективности всего процесса обработки данных, вычислительная система может производить объединение альтернативных семантико-синтаксических структур, полученных в ходе семантико-синтаксического анализа, в одну итоговую структуру. Далее вычислительная система может проводить интерпретацию множества полученных

семантических структур, используя набор продукционных правил для извлечения информационных объектов (к примеру, именованных сущностей). Распознавание именованных сущностей (Named-entity recognition, или NER), также известное как идентификация сущностей, выявление сущностей и извлечение сущностей, представляет собой задачу извлечения информации, результатом которой служит выявление токенов в тексте на естественном языке и их классификация по заранее определенным категориям - таким, к примеру, как имена персон, названия организаций, адреса или географические координаты, представление времени, количества, денежные единицы, проценты и т.д. Категории именованных сущностей и (или) семантические классы, отвечающие иным информационным объектам, извлекаемым из текста на естественном языке, могут быть представлены классами онтологии - как предопределенной, так и динамически выстраиваемой.

[00026] Термин «онтология» в контексте настоящего документа означает модель, представляющую объекты, относящиеся к определенной области знаний (предметной области), а также отношения между такими объектами. Онтология может содержать определения множества классов (концептов). Определение класса может ссылаться на один или более экземпляров концепта, т.е. информационных объектов. Информационный объект может представлять собой объект реального мира (к примеру, персону или предмет) и (или) некоторые характеристики, связанные с одним или более объектами реального мира (к примеру, измеримый атрибут или некоторое качество).

[00027] Продукционные правила, используемые для интерпретации семантических структур, могут включать правила интерпретации и правила идентификации. Правило интерпретации может содержать левую часть, представленную набором логических выражений, определенных на одном или более шаблонах семантических структур, и правую часть, представленную одним или более утверждениями относительно информационных объектов, представляющих сущности, на которые имеется ссылка в тексте на естественном языке.

[00028] В результате наложения шаблона, определяемого левой частью продукционного правила, на семантическую структуру, представляющую по меньшей мере часть предложения в тексте на естественном языке, может быть приведена в действие правая часть продукционного правила. Правая часть продукционного правила может устанавливать ассоциативную связь между одним или более атрибутами (отражающими лексические, синтаксические и (или) семантические свойства слов из первоначального предложения) и информационными объектами, представленными узлами. В одном из иллюстративных примеров правая часть правила интерпретации может представлять собой утверждение, устанавливающее ассоциативную связь между токеном из текста на естественном языке и категорией именованных сущностей.

[00029] Правило идентификации может использоваться для установления ассоциативной связи для пары информационных объектов, которые представляют одну и ту же сущность из 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.

[00030] Поскольку альтернативные семантико-синтаксические структуры, соответствующие одному и тому же фрагменту (к примеру, одному и тому же предложению) текста на естественном языке, являются взаимоисключающими, такими

же являются и информационные объекты, которые могут быть извлечены из таких семантико-синтаксических структур. Таким образом, в каждой группе найденных альтернативных объектов следует выбрать один информационный объект. В некоторых вариантах реализации изобретения вычислительная система может выбирать из группы найденных альтернативных объектов один информационный объект, извлеченный при помощи семантико-синтаксической структуры, характеризующейся оптимальным (к примеру, максимальным или минимальным) значением метрики качества среди множества семантико-синтаксических структур, как подробнее описано ниже в настоящем документе.

[00031] Далее вычислительная система может применить один или более методов извлечения фактов для выявления в тексте на естественном языке одного или более фактов, ассоциирующихся с определенными информационными объектами. Термин «факт» в контексте настоящего документа означает отношение между информационными объектами, на которые имеется ссылка в тексте на естественном языке. Примерами таких отношений могут быть работа лица X по найму в организационном подразделении Y, расположение объекта X в географической точке Y, приобретение организационной единицы X организационной единицей Y и т.д. Таким образом, факт может быть связан с одной или более категориями фактов и/или сущностей. К примеру, факт, связанный с неким лицом, может относиться к дате его рождения, образованию, роду занятий, месту работы и т.д. В другом примере факт, связанный с коммерческой сделкой, может относиться к типу сделки и к сторонам этой сделки, к обязательствам сторон, дате подписания договора, дате совершения сделки, расчетам по договору и т.д. Извлечение фактов предполагает выявление различных отношений между извлеченными информационными объектами.

[00032] В некоторых вариантах реализации изобретения извлечение фактов может предусматривать интерпретацию множества семантических структур с использованием набора продукционных правил, в том числе правил интерпретации и (или) правил идентификации, как подробнее описано ниже в настоящем документе. Поскольку объединяемые альтернативные семантико-синтаксические структуры, соответствующие одному и тому же фрагменту (к примеру, одному и тому же предложению) текста на естественном языке, являются взаимоисключающими, такими же являются и факты, которые могут быть извлечены из таких семантико-синтаксических структур. Таким образом, в каждой группе найденных альтернативных фактов следует выбрать один факт. В некоторых вариантах реализации изобретения вычислительная система может выбирать из группы найденных фактов один факт, извлеченный из семантико-синтаксической структуры, характеризующейся оптимальным (к примеру, максимальным или минимальным) значением метрики качества среди множества семантико-синтаксических структур, как подробнее описано ниже в настоящем документе.

[00033] Системы и способы, представленные в настоящем документе, могут быть реализованы аппаратно (например, с помощью универсальных и (или) специализированных устройств обработки и (или) иных устройств и соответствующих электронных схем), программно (например, с помощью команд, выполняемых устройством обработки) или сочетанием этих подходов. Различные варианты реализации упомянутых выше способов и систем подробно описаны ниже в этом документе на примерах, без каких бы то ни было ограничений.

[00034] На Фиг. 1 изображена блок-схема примера реализации способа извлечения информации с использованием альтернативных вариантов семантико-синтаксического разбора в соответствии с одним или более вариантами реализации настоящего

изобретения. Способ 100 и (или) каждая из его отдельно взятых функций, процедур, подпрограмм и каждое из действий могут осуществляться с помощью одного или более процессоров вычислительной системы (к примеру, вычислительной системы 100 на Фиг. 1), реализующей этот способ. В некоторых вариантах реализации способ 100 может осуществляться в одном потоке обработки. При альтернативном подходе способ 100 может осуществляться с использованием двух или более потоков обработки, при этом в каждом потоке реализована(о) одна (одно) или несколько 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, в которых реализован способ 100, могут выполняться асинхронно по отношению друг к другу. Таким образом, несмотря на то что Фиг. 1 и соответствующее описание содержат перечень действий для способа 100 в определенном порядке, в различных вариантах осуществления способа по меньшей мере некоторые из описанных операций могут выполняться параллельно и (или) в случайно выбранном порядке.

[00035] На шаге 110 блок-схемы вычислительная система, реализующая способ 100, может выполнить семантико-синтаксический анализ исходного текста 101 на естественном языке, который может быть представлен, к примеру, одним или более исходными документами. В результате семантико-синтаксического анализа может получаться множество семантических структур, представляющих предложения в тексте на естественном языке. Каждая семантическая структура может быть представлена ациклическим графом, который включает множество узлов, соответствующих семантическим классам, и множество ребер, соответствующих семантическим отношениям. Ради простоты любое подмножество семантической структуры в этом документе мы будем называть «структурой» (а не «подструктурой»), если только предметом рассмотрения не является отношение типа «родительский элемент - дочерний элемент» (предок-потомок) между двумя семантическими структурами.

[00036] Семантико-синтаксический анализ исходного текста на естественном языке может предусматривать выполнение для каждого предложения лексико-морфологического анализа, а затем грубого синтаксического анализа и обработка полученных синтаксических деревьев с целью получения семантико-синтаксической структуры, соответствующей предложению, как подробнее описано ниже в настоящем документе со ссылкой на Фиг. 2-12.

[00037] На нескольких этапах семантико-синтаксического анализа могут получаться множественные альтернативные результаты - как промежуточные, так и финальные. К примеру, ввиду омонимии и (или) совпадения грамматических форм, соответствующих разным лексико-морфологическим значениям определенного слова, для данного слова в ходе лексико-морфологического анализа может быть установлено множество морфологических значений. Кроме того, в ходе грубого синтаксического анализа к одному предложению может быть использовано множество синтаксических моделей, применимых к данному предложению, что может приводить ко множественным версиям итогового синтаксического дерева. Таким образом, результатом точного синтаксического анализа могут быть множество синтаксических деревьев, соответствующих заданному первоначальному предложению.

[00038] Хотя известны способы и приемы, использующие функции метрики качества, с учетом совместимости лексических значений слов в первоначальном предложении, поверхностных отношений, глубинных отношений и т.д., с целью выбора наилучшего

из синтаксических деревьев, соответствующих исходному предложению, средства и способы, изложенные в настоящем описании изобретения, расширяют функциональность универсальных или специализированных вычислительных устройств возможностью достижения более полного извлечения информации за счет учета множественности синтаксических деревьев.

[00039] На шаге 120 блок-схемы вычислительная система может производить слияние альтернативных семантико-синтаксических структур, полученных в ходе семантико-синтаксического анализа, в одну объединенную структуру. Процедура слияния может предусматривать объединение в один граф узлов и ребер графов, представляющих альтернативные семантико-синтаксические структуры.

[00040] Каждой из семантико-синтаксических структур может соответствовать одно из значений метрики качества. Метрика качества может учитывать совместимость лексических значений слов в первоначальном предложении, поверхностных отношений, глубинных отношений и т.д. В некоторых вариантах реализации изобретения численные значения метрики качества используются для выбора информационного объекта из множества объектов-претендентов, извлеченных при помощи альтернативных семантико-синтаксических структур, как подробнее описано ниже в настоящем документе.

[00041] В некоторых вариантах реализации изобретения процедура слияния может предусматривать обнаружение дублирующих подструктур, цель которого - либо недопущение в дальнейшем присутствия дубликатов одной и той же подструктуры в полученной структуре, либо удаление таких дубликатов из полученной структуры.

[00042] На шаге 130 блок-схемы вычислительная система может проводить интерпретацию множества полученных семантических структур, используя набор продукционных правил для извлечения множества информационных объектов (к примеру, именованных сущностей). Категории именованных сущностей и (или) семантические классы, отвечающие иным информационным объектам, извлекаемым из текста на естественном языке, могут быть представлены концептами онтологии - как предопределенной, так и динамически выстраиваемой.

[00043] В некоторых вариантах реализации изобретения вычислительная система может применять множественные альтернативные наборы продукционных правил к одним и тем же семантико-синтаксическим структурам. Поскольку альтернативные наборы продукционных правил являются взаимоисключающими, такими же являются и информационные объекты, которые могут быть извлечены при наложении таких наборов продукционных правил. Таким образом, в каждой группе найденных альтернативных объектов следует выбрать один информационный объект. В некоторых вариантах реализации изобретения вычислительная система может выбирать из группы найденных альтернативных объектов один информационный объект, извлеченный при помощи набора правил, характеризующегося максимальным значением веса среди альтернативных наборов правил.

[00044] Продукционные правила, используемые для интерпретации семантических структур, могут представлять собой правила интерпретации и правила идентификации. Правило интерпретации может содержать левую часть, представленную набором 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.

[00045] Шаблон семантической структуры может содержать некоторые элементы

семантической структуры (например, принадлежность к определенному лексическому/семантическому классу, нахождение в некоторой поверхностной или глубинной позиции, наличие определенной граммемы или семантемы 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 на семантико-синтаксическом дереве. В одном из иллюстративных примеров такая операция может проверять один из узлов на принадлежность к поддереву другого узла.

[00046] В результате наложения шаблона, определяемого левой частью продукционного правила, на семантическую структуру, представляющую по меньшей мере часть предложения в тексте на естественном языке, может быть приведена в действие правая часть продукционного правила. Правая часть продукционного правила может устанавливать ассоциативную связь между одним или более атрибутами (отражающими лексические, синтаксические и (или) семантические свойства слов из первоначального предложения) и информационными объектами, представленными узлами. В одном из иллюстративных примеров правая часть правила интерпретации может представлять собой утверждение, устанавливающее ассоциативную связь между токеном из текста на естественном языке и категорией именованных сущностей.

[00047] Правило идентификации может использоваться для установления ассоциативной связи для пары информационных объектов, которые представляют одну и ту же сущность из 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. В некоторых вариантах реализации изобретения вычислительная система может повышать на некоторую предопределенную или динамически определяемую величину значение метрики качества информационного объекта, для которого установлено, что он ссылается на ту же сущность из реального мира, что и другой информационный объект, характеризующийся более высоким начальным значением метрики качества, как подробнее описано ниже в настоящем документе.

[00048] Несмотря на то что в иллюстративном примере на Фиг. 1 извлечение информационных объектов производится путем интерпретации множества семантических структур при помощи набора продукционных правил, в различных альтернативных вариантах реализации изобретения могут использоваться функции классификатора, в которых могут, наряду с лексическими и морфологическими признаками, использовать синтаксические и (или) семантические признаки, полученные при семантико-синтаксическом анализе текста на естественном языке. В некоторых вариантах реализации изобретения всевозможные лексические, грамматические и (или) семантические атрибуты токена естественного языка могут использоваться в составе одной или более функций классификатора. Каждая функция классификатора может определять для токена естественного языка степень ассоциативной связи с определенной категорией информационных объектов.

[00049] В одном из иллюстративных примеров функция классификатора может обучаться на обучающей выборке из текстов на естественном языке, которые были размечены с помощью систем и способов согласно одному или более вариантам реализации настоящего изобретения. При построении функции классификатора в ней

могут быть реализованы различные методы - от наивного байесовского классификатора до техники дифференциальной эволюции, метода опорных векторов, алгоритмов случайного леса, нейронных сетей, градиентного бустинга и т.д. Обучение классификатора может включать определение наиболее важных атрибутов текстов на естественном языке и (или) настройку значений одного или более параметров функции классификатора. После завершения стадии обучения функция классификатора может использоваться для обработки подтверждающего набора текстов на естественном языке (т.е. неразмеченных текстов). Качество классификации можно оценить, применяя классификатор к одному или более размеченным текстам на естественном языке из тестового набора. Обученная функция классификатора может использоваться для получения значения, отражающего степень ассоциативной связи определенной части текста на естественном языке с определенной категорией фактов, информационных объектов, других текстов на естественном языке и т.д.

[00050] В некоторых вариантах реализации изобретения способ извлечения информационных объектов может предусматривать использование продукционных правил в сочетании с моделями классификаторов.

[00051] Как отмечалось выше в тексте настоящего документа, продукционные правила и (или) функции классификатора применяются к объединенной семантико-синтаксической структуре, в состав которой могут входить альтернативные семантико-синтаксические подструктуры, полученные в ходе семантико-синтаксического анализа. В результате из одного и того же предложения (или одной и той же части предложения) в тексте на естественном языке могут быть извлечены множественные информационные объекты. Поскольку объединенные альтернативные семантико-синтаксические структуры, соответствующие одному и тому же фрагменту (к примеру, одному и тому же предложению) текста на естественном языке, являются взаимоисключающими, такими же являются и информационные объекты, которые могут быть извлечены из таких семантико-синтаксических структур. Таким образом, в каждой группе найденных альтернативных объектов следует выбрать один информационный объект. В некоторых вариантах реализации изобретения вычислительная система может выбирать из группы найденных альтернативных объектов один информационный объект, извлеченный из семантико-синтаксической структуры, характеризующейся оптимальным значением метрики качества среди альтернативных семантико-синтаксических структур. В одном из иллюстративных примеров функция метрики качества может учитывать совместимость лексических значений слов в первоначальном предложении, поверхностных отношений, глубинных отношений и (или) иных всевозможных параметров каждой семантико-синтаксической структуры.

[00052] В некоторых вариантах реализации изобретения вычислительная система после извлечения информационных объектов из фрагмента текста на естественном языке может разрешать кореференциальные и анафорические ссылки между токенами текста на естественном языке, ассоциированными с извлеченными информационными объектами. Термин «кореференция» в контексте настоящего документа означает конструкцию естественного языка, содержащую два или более токенов естественного языка, которые относятся к одной сущности (например, к одному и тому же лицу, предмету, месту, организации и т.д.).

[00053] После разрешения кореференций вычислительная система может присвоить альтернативным информационным объектам, извлеченным из одного и того же фрагмента текста на естественном языке, начальные значения рейтинга на основе значений метрики качества соответствующих семантико-синтаксических структур,

использованных в ходе извлечения информации. После этого вычислительная система может повышать на некоторую предопределенную или динамически определяемую величину значение рейтинга информационного объекта, для которого установлено наличие кореференций на ту же сущность из реального мира, что и другой информационный объект, характеризующийся более высоким начальным значением рейтинга. После этого вычислительная система может выбрать информационный объект, связанный ассоциативной связью с оптимальным значением метрики качества среди альтернативных информационных объектов.

[00054] На шаге 140 блок-схемы вычислительная система может применить один или более методов извлечения фактов для выявления в тексте на естественном языке одного или более фактов, ассоциированных с определенными информационными объектами. Термин «факт» в контексте настоящего документа означает отношение между информационными объектами, на которые имеется ссылка в тексте на естественном языке. Примерами таких отношений могут быть работа лица X по найму в организационном подразделении Y, расположение объекта X в географической точке Y, приобретение организационной единицы X организационной единицей Y и т.д. Таким образом, факт может быть связан ассоциативной связью с одной или более категориями фактов. К примеру, факт, связанный с неким лицом, может иметь отношение к дате его рождения, образованию, роду занятий, месту работы и т.д. В другом примере факт, связанный с коммерческой сделкой, может иметь отношение к типу сделки и к сторонам этой сделки, к обязательствам сторон, дате подписания договора, дате совершения сделки, расчетам по договору и т.д. Извлечение фактов предполагает выявление различных отношений между извлеченными информационными объектами.

[00055] В некоторых вариантах реализации изобретения извлечение фактов может предусматривать интерпретацию множества семантических структур с использованием набора продукционных правил, в том числе правил интерпретации и (или) правил идентификации, как подробнее описано ниже в настоящем документе. В дополнение к этому или в качестве альтернативы извлечение фактов может предусматривать использование одной или более функций классификатора для обработки всевозможных лексических, грамматических и (или) семантических атрибутов предложения на естественном языке. Каждая функция классификатора может определять степень ассоциативной связи по меньшей мере части предложения на естественном языке с определенной категорией фактов.

[00056] Как отмечалось выше в тексте настоящего документа, продукционные правила и (или) функции классификатора применяются к объединенной семантико-синтаксической структуре, в состав которой могут входить альтернативные семантико-синтаксические подструктуры, полученные в ходе семантико-синтаксического анализа. В результате из одного и того же предложения (или одной и той же части предложения) в тексте на естественном языке могут быть извлечены альтернативные отношения между одними и теми же парами или группами информационных объектов. Поскольку объединенные альтернативные семантико-синтаксические структуры, соответствующие одному и тому же фрагменту (к примеру, одному и тому же предложению) текста на естественном языке, являются взаимоисключающими, такими же являются и факты, которые могут быть извлечены из таких семантико-синтаксических структур. Таким образом, в каждой группе найденных альтернативных фактов следует выбрать один факт. В некоторых вариантах реализации изобретения вычислительная система может выбирать из группы найденных альтернативных фактов один факт, извлеченный при помощи семантико-синтаксической структуры, характеризующейся оптимальным

значением метрики качества среди множественных альтернативных семантико-синтаксических структур. В одном из иллюстративных примеров функция метрики качества может учитывать совместимость лексических значений слов в первоначальном предложении, поверхностных отношений, глубинных отношений и (или) иных всевозможных параметров каждой семантико-синтаксической структуры.

[00057] В некоторых вариантах реализации изобретения извлечение некоторых информационных объектов и их отношений (фактов) может осуществляться на основе других информационных объектов (к примеру, информационный объект, у которого имеется ссылка на родной город некоего лица, извлекается на основе информационного объекта, у которого имеется ссылка на это лицо). Если некоторый информационный объект был отсеян в силу того, что одна или более альтернативных семантико-синтаксических структур были отсеяны по критерию неоптимальности значения метрики качества, то вычислительная система может приступить к удалению из полученного набора данных информационных объектов и отношений, производных от уже отсеянного информационного объекта.

[00058] В некоторых вариантах реализации изобретения вычислительная система может представлять информационные объекты и их отношения в виде графа RDF. RDF (Resource Definition Framework - среда описания 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 задавать ассоциативную связь между токеном из текста на естественном языке и категорией именованных сущностей.

[00059] На шаге 150 блок-схемы вычислительная система может использовать извлеченные информационные объекты и факты для выполнения самых разных задач обработки текстов на естественном языке - к примеру, задач машинного перевода, семантического поиска, классификации документов, кластеризации, фильтрации текста и т.д. В том случае, если окажутся выполнены все действия, описанные с отсылкой к шагу 150 блок-схемы, способ может завершиться.

[00060] На Фиг. 2 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 200 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 200 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 vim. д.

[00061] 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 приведено ниже в настоящем документе со ссылкой на Фиг. 3.

[00062] 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.

[00063] 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.

[00064] 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).

[00065] Фиг. 3 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" 312 и "will" 314. Грамматическим значением, соотнесенным с лексическим значением 312, является <Verb, GTVerbModal, ZeroType, Present, Nonnegative, Composite II>. Грамматическим значением, соотнесенным с лексическим значением 314, является <Verb, GTVerbModal, ZeroType, Present, Nonnegative, Irregular, Composite II>.

[00066] Фиг. 4 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 210 представляет собой модель определенного естественного языка.

[00067] 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.

[00068] Фиг. 5 схематически иллюстрирует несколько примеров морфологических описаний. В число компонентов морфологических описаний 201 могут входить: описания словоизменения 310, грамматическая система 320, описания словообразования 330 и другие. Грамматическая система 320 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 310 и описания словообразования 330.

[00069] Описание словоизменения 310 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 330 определяет, какие новые слова могут быть образованы от данного слова (например, сложные слова).

[00070] 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.

[00071] На Фиг. 6 приведены примеры синтаксических описаний. В число компонентов синтаксических описаний 202 могут входить, среди прочего, поверхностные модели 410, описания поверхностных позиций 420, описание референциального и структурного контроля 456, описание управления и согласования 440, описание недревесного синтаксиса 450 и правила анализа 460. 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.

[00072] Поверхностные модели 410 могут быть представлены в виде совокупностей одной или нескольких синтаксических форм («синтформ» 412) 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 410. 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.

[00073] В модели составляющих может использоваться множество поверхностных позиций 415 дочерних составляющих и описаний их линейного порядка 416 для описания грамматических значений 414 возможных заполнителей этих поверхностных позиций. Диатезы 417 представляют собой соответствия между поверхностными позициями 415 и глубинными позициями 514 (как показано на Фиг. 8). Коммуникативные описания 480 описывают коммуникативный порядок в предложении.

[00074] Описание линейного порядка (416) может быть представлено в виде выражений линейного порядка, отражающих последовательность, в которой различные поверхностные позиции (415) 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 415, соответствующих порядку слов.

[00075] Коммуникативные описания 480 могут описывать порядок слов в синтформе 412 с точки зрения коммуникативных актов, представленных в виде коммуникативных выражений порядка, которые похожи на выражения линейного порядка. Описания управления и согласования 440 может включать правила и ограничения на грамматические значения присоединяемых составляющих, которые используются во время синтаксического анализа.

[00076] Описания недревесного синтаксиса 450 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 не древесного синтаксиса 450 могут, среди прочего, включать описание эллипсиса 452, описания согласования 454, а также описания референциального и структурного контроля 430.

[00077] Правила анализа 460 могут описывать свойства конкретного языка и использоваться в рамках семантического анализа. Правила анализа 460 могут включать правила вычисления семантем 462 и правила нормализации 464. Правила нормализации 464 могут использоваться для описания трансформаций семантических структур, которые могут отличаться в разных языках.

[00078] На Фиг. 7 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.

[00079] 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.

[00080] 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.

[00081] 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.

[00082] Система семантем 530 может представлять собой множество семантических категорий и семантем, которые представляют значения семантических категорий. В качестве иллюстративного примера семантическая категория "DegreeOfComparison" (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.

[00083] 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.

[00084] 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.

[00085] 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.

[00086] На Фиг. 8 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.

[00087] Лексическое значение 612 в лексико-семантической иерархии 510 может быть соотнесено с поверхностной моделью 410, которая в свою очередь через одну или несколько диатез 417 может быть соотнесена с соответствующей глубинной моделью 512. Лексическое значение 612 может наследовать семантический класс своего родителя и уточнять свою глубинную модель 512.

[00088] Поверхностная модель 410 лексического значения может включать одну или несколько синтаксических форм 412. Синтформа 412 поверхностной модели 410 может включать одну или несколько поверхностных позиций 415, в том числе соответствующие описания их линейного порядка 416, одно или несколько грамматических значений 414, выраженных в виде набора грамматических категорий (граммем), одно или несколько семантических ограничений, соотнесенных с заполнителями поверхностных позиций, и одну или несколько диатез 417. 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.

[00089] Фиг. 9 схематически иллюстрирует примеры структур данных, которые могут быть использованы в рамках одного или более методов настоящего изобретения. Снова ссылаясь на Фиг. 2, в блоке 214 вычислительное устройство, реализующее данный способ, может проводить лексико-морфологический анализ предложения 212 для построения лексико-морфологической структуры 722 согласно Фиг. 9. Лексико-морфологическая структура 722 может включать множество соответствий лексического

и грамматического значений для каждой лексической единицы (например, слова) исходного предложения. Фиг. 3 схематически иллюстрирует пример лексико-морфологической структуры.

[00090] Снова возвращаясь к Фиг. 2, 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.

[00091] 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.

[00092] 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 415 множества родительских составляющих, с тем чтобы были охвачены все лексические единицы исходного предложения 212.

[00093] 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 414, например, на основе частей речи и отношений между ними. Фиг. 10 схематически иллюстрирует пример графа обобщенных составляющих.

[00094] В блоке 216 вычислительное устройство может проводить точный синтаксический анализ предложения 212 для формирования одного или более синтаксических деревьев 742 согласно Фиг. 9 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.

[00095] 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.

[00096] 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 460 и предполагает соотнесение одного или более атрибутов (отражающих лексические, синтаксические и (или) семантические свойства слов исходного предложения 212) с каждым семантическим классом.

[00097] На Фиг. 11 приводится пример синтаксической структуры предложения, сгенерированной из графа обобщенных составляющих, показанного на Фиг. 10 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).

[00098] На Фиг. 12 приводится семантическая структура, соответствующая синтаксической структуре на Фиг. 11. В отношении вышеупомянутого лексического элемента "life" (жизнь) (906) на Фиг. 12 семантическая структура включает лексический класс 1010 и семантический класс 1030, соответствующие представленным на Фиг. 11, однако вместо поверхностной позиции (905) семантическая структура включает глубинную позицию "Sphere" (сфера_деятельности) 1020.

[00099] 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 означать концепт, а принадлежащий классу объект может означать экземпляр данного концепта.

[000100] 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.

[000101] 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.

[000102] 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.

[000103] 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.

[000104] 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.

[000105] На Фиг. 13 схематически показан иллюстративный пример вычислительного устройства (1000), которое может исполнять набор команд, которые вызывают выполнение вычислительным устройством любого отдельно взятого или нескольких способов настоящего изобретения. Вычислительное устройство может подключаться к другому вычислительному устройству по локальной сети, корпоративной сети, сети экстранет или сети Интернет. Вычислительное устройство может работать в качестве сервера или клиентского вычислительного устройства в сетевой среде "клиент/сервер" либо в качестве однорангового вычислительного устройства в одноранговой (или распределенной) сетевой среде. Вычислительное устройство может быть представлено персональным компьютером (ПК), планшетным ПК, телевизионной приставкой (STB), карманным ПК (PDA), сотовым телефоном или любым вычислительным устройством, способным выполнять набор команд (последовательно или иным образом), определяющих операции, которые должны быть выполнены этим вычислительным устройством. Кроме того, в то время как показано только одно вычислительное устройство, следует принять, что термин «вычислительное устройство» также может включать любую совокупность вычислительных устройств, которые отдельно или совместно выполняют набор (или несколько наборов) команд для выполнения одной или нескольких методик, описанных в настоящем документе.

[000106] Пример вычислительного устройства (1000) включает процессор (502), основную память (504) (например, постоянное запоминающее устройство (ПЗУ) или динамическую оперативную память (DRAM)) и устройство хранения данных (518), которые взаимодействуют друг с другом по шине (530).

[000107] Процессор (502) может быть представлен одним или более универсальными вычислительными устройствами, 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 одно или несколько вычислительных устройств специального назначения, например, заказную интегральную микросхему (ASIC), программируемую пользователем вентильную матрицу (FPGA), процессор цифровых сигналов (DSP), сетевой процессор и т.п. Процессор (502) настроен на выполнение команд (526) для осуществления рассмотренных в настоящем документе операций и функций.

[000108] Вычислительное устройство (1000) может дополнительно включать устройство сетевого интерфейса (522), устройство визуального отображения (510), устройство ввода символов (512) (например, клавиатуру), и устройство ввода - сенсорный экран (514).

[000109] Устройство хранения данных (518) может содержать машиночитаемый носитель данных (524), в котором хранится один или более наборов команд (526), и в котором реализован один или более из методов или функций настоящего изобретения. Команды (526) также могут находиться полностью или по меньшей мере частично в

основной памяти (504) и/или в процессоре (502) во время выполнения их в вычислительном устройстве (1000), при этом оперативная память (504) и процессор (502) также составляют машиночитаемый носитель данных. Команды (526) дополнительно могут передаваться или приниматься по сети (516) через устройство сетевого интерфейса (522).

[000110] В некоторых вариантах реализации изобретения набор команд 526 может содержать команды способа 100 извлечения информации с использованием альтернативных вариантов семантико-синтаксического разбора в соответствии с одним или более вариантами реализации настоящего изобретения. Хотя машиночитаемый носитель данных 524 показан в примере на Фиг. 13 в виде одного носителя, термин «машиночитаемый носитель» следует понимать в широком смысле, подразумевающем один носитель или более носителей (к примеру, централизованную или распределенную базу данных и (или) соответствующие кэши и серверы), в которых хранится один или более наборов команд. Кроме того, термин «машиночитаемый носитель данных» следует понимать в широком смысле, подразумевающем любой носитель, способный хранить, кодировать или переносить набор команд для выполнения вычислительной машиной и обеспечивающий реализацию на вычислительной машине любой одной или более методик настоящего изобретения. Поэтому термин «машиночитаемый носитель данных» относится, помимо прочего, к твердотельной памяти, а также к оптическим и магнитным носителям.

[000111] 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 - к примеру, 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.

[000112] 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.

[000113] Некоторые части описания предпочтительных вариантов реализации изобретения представлены в виде алгоритмов и символического представления операций с битами данных в памяти компьютера. Такие описания и представления алгоритмов представляют собой средства, используемые специалистами в области обработки данных, что обеспечивает наиболее эффективную передачу сути 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 не обязательно, 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.

[000114] 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 не указано дополнительно, принимается, что в последующем описании термины «определение», «вычисление», «расчет», «получение», «установление», «выявление», «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 устройствах памяти вычислительной системы, в другие данные, также представленные в виде физических величин в устройствах памяти или реестрах вычислительной системы или иных устройствах хранения, передачи или отображения такой информации.

[000115] 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 перенастраивается с помощью программы, хранящейся в памяти компьютера. Такая компьютерная программа может храниться на машиночитаемом носителе данных - к примеру, помимо всего 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.

[000116] 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:

выполнение вычислительной системой семантико-синтаксического анализа по меньшей мере одной части текста на естественном языке с целью получения множества семантико-синтаксических структур, представляющих по меньшей мере одну указанную часть текста на естественном языке, при этом множество семантико-синтаксических структур включает первую альтернативную семантико-синтаксическую структуру и вторую альтернативную семантико-синтаксическую структуру;

объединение множества семантико-синтаксических структур с целью получения объединенной семантико-синтаксической структуры;

исключение дублирующих семантико-синтаксических подструктур из объединенной семантико-синтаксической структуры;

выявление в пределах указанной по меньшей мере одной части текста на естественном языке одного или более информационных объектов путем интерпретации объединенной семантико-синтаксической структуры с целью установления ассоциативной связи одного или более токенов, образованных указанной по меньшей мере одной частью текста на естественном языке, с по меньшей мере одной категорией информационных объектов,

при этом указанная интерпретация объединенной семантико-синтаксической структуры производится с учетом значения метрики качества, ассоциированной с по меньшей мере частью первой альтернативной семантико-синтаксической структуры по меньшей мере одной части текста на естественном языке.

2. Способ по п. 1, в котором интерпретация объединенной семантико-синтаксической структуры определяет для одного или более токена степень ассоциативной связи с определенной категорией информационных объектов.

3. Способ по п. 1, в котором каждая семантико-синтаксическая структура из множества семантико-синтаксических структур представлена графом, который включает множество узлов, соответствующих множеству семантических классов, и множество ребер, соответствующих множеству семантических отношений.

4. Способ по п. 1, в котором интерпретация объединенной семантико-синтаксической структуры дополнительно предусматривает:

применение к объединенной семантико-синтаксической структуре набора продукционных правил.

5. Способ по п. 1, в котором интерпретация объединенной семантико-синтаксической структуры дополнительно предусматривает:

применение функции классификатора к одному или более значениям лексических, и/или грамматических, и/или синтаксических, и/или семантических атрибутов.

6. Способ по п. 1, дополнительно предусматривающий:

выявление одного или более отношений между найденными информационными объектами с целью извлечения одного или более фактов, отраженных в по меньшей мере одной части текста на естественном языке.

7. Способ по п. 6, в котором выявление одного или более отношений дополнительно предусматривает:

интерпретацию семантико-синтаксических структур с использованием набора продукционных правил.

8. Способ по п. 6, в котором выявление одного или более отношений дополнительно предусматривает:

применение функции классификатора к одному или более значениям лексических, и/или грамматических, и/или синтаксических, и/или семантических атрибутов.

9. Способ по п. 1, дополнительно включающий:

выбор из двух или более информационных объектов, извлеченных из одного и того же фрагмента текста на естественном языке, по меньшей мере одного информационного объекта с оптимальным значением метрики качества.

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

запоминающее устройство (ЗУ);

процессор, связанный с указанным ЗУ, причем процессор настроен на:

выполнение семантико-синтаксического анализа по меньшей мере одной части текста на естественном языке с целью получения множества семантико-синтаксических структур, представляющих по меньшей мере одну указанную часть текста на естественном языке, при этом множество семантико-синтаксических структур включает первую альтернативную семантико-синтаксическую структуру и вторую альтернативную семантико-синтаксическую структуру;

объединение множества семантико-синтаксических структур с целью получения объединенной семантико-синтаксической структуры;

исключение дублирующих семантико-синтаксических подструктур из объединенной семантико-синтаксической структуры;

выявление в пределах указанной по меньшей мере одной части текста на естественном языке одного или более информационных объектов путем интерпретации объединенной семантико-синтаксической структуры с целью установления ассоциативной связи одного или более токенов, образованных указанной по меньшей мере одной частью текста на естественном языке, с по меньшей мере одной категорией информационных объектов, при этом указанная интерпретация объединенной семантико-синтаксической структуры производится с учетом значения метрики качества, ассоциированной с по меньшей мере частью первой альтернативной семантико-синтаксической структуры по меньшей мере одной части текста на естественном языке.

11. Система по п. 10, в которой интерпретация объединенной семантико-синтаксической структуры определяет для одного или более токена степень ассоциативной связи с определенной категорией информационных объектов.

12. Система по п. 10, в которой каждая семантико-синтаксическая структура из множества семантико-синтаксических структур представлена графом, который включает множество узлов, соответствующих множеству семантических классов, и множество ребер, соответствующих множеству семантических отношений.

13. Система по п. 10, в которой интерпретация объединенной семантико-синтаксической структуры дополнительно предусматривает:

применение к объединенной семантико-синтаксической структуре набора продукционных правил.

14. Система по п. 10, в которой интерпретация объединенной семантико-синтаксической структуры дополнительно предусматривает:

применение функции классификатора к одному или более значениям лексических, грамматических, синтаксических или семантических атрибутов.

15. Система по п. 10, в которой процессор дополнительно настроен на:

выявление одного или более отношений между найденными информационными объектами с целью извлечения одного или более фактов, отраженных по меньшей мере в одной части текста на естественном языке.

16. Энергонезависимый машиночитаемый носитель данных, содержащий исполняемые команды для извлечения информации из текстов на естественном языке, которые при их выполнении вычислительной системой обеспечивают выполнение указанной системой таких операций, как:

выполнение семантико-синтаксического анализа по меньшей мере одной части текста на естественном языке с целью получения множества семантико-синтаксических структур, представляющих по меньшей мере одну указанную часть текста на естественном языке, при этом множество семантико-синтаксических структур включает первую альтернативную семантико-синтаксическую структуру и вторую альтернативную семантико-синтаксическую структуру;

объединение множества семантико-синтаксических структур с целью получения объединенной семантико-синтаксической структуры;

исключение дублирующих семантико-синтаксических подструктур из объединенной семантико-синтаксической структуры;

выявление в пределах указанной по меньшей мере одной части текста на естественном языке одного или более информационных объектов путем интерпретации объединенной семантико-синтаксической структуры с целью установления ассоциативной связи одного или более токенов, образованных указанной по меньшей мере одной частью текста на естественном языке, с по меньшей мере одной категорией информационных объектов, при этом указанная интерпретация объединенной семантико-синтаксической структуры

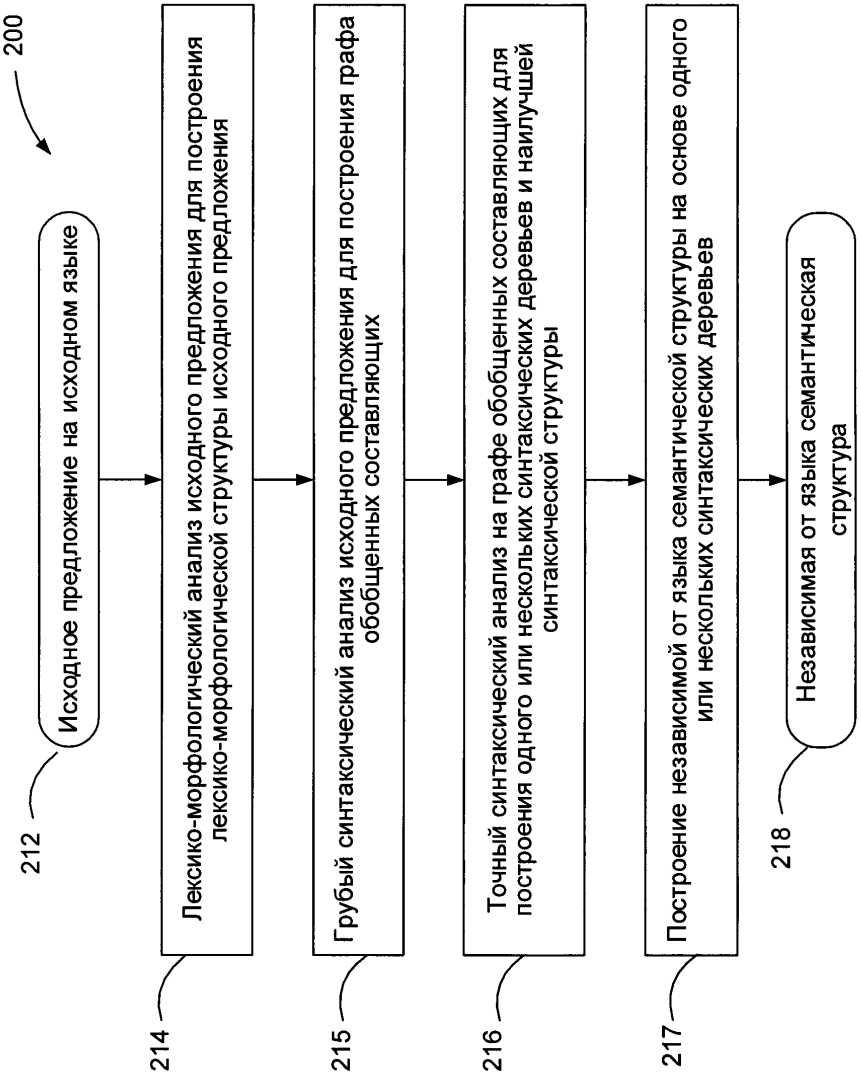
производится с учетом значения метрики качества, ассоциированной с по меньшей мере частью первой альтернативной семантико-синтаксической структуры по меньшей мере одной части текста на естественном языке.

1


100
101
Текст на естественном языке
110
Семантико-синтаксически анализ
120
Слияние альтернативных семантико-синтаксических стркутур
130
Извлечение информационного объекта на базе объединенной семантико-синтаксической структуры
140
Извлечение факта на базе объединенной семантико-синтаксической структуры
150
Выполнение задач обработки естественного языка

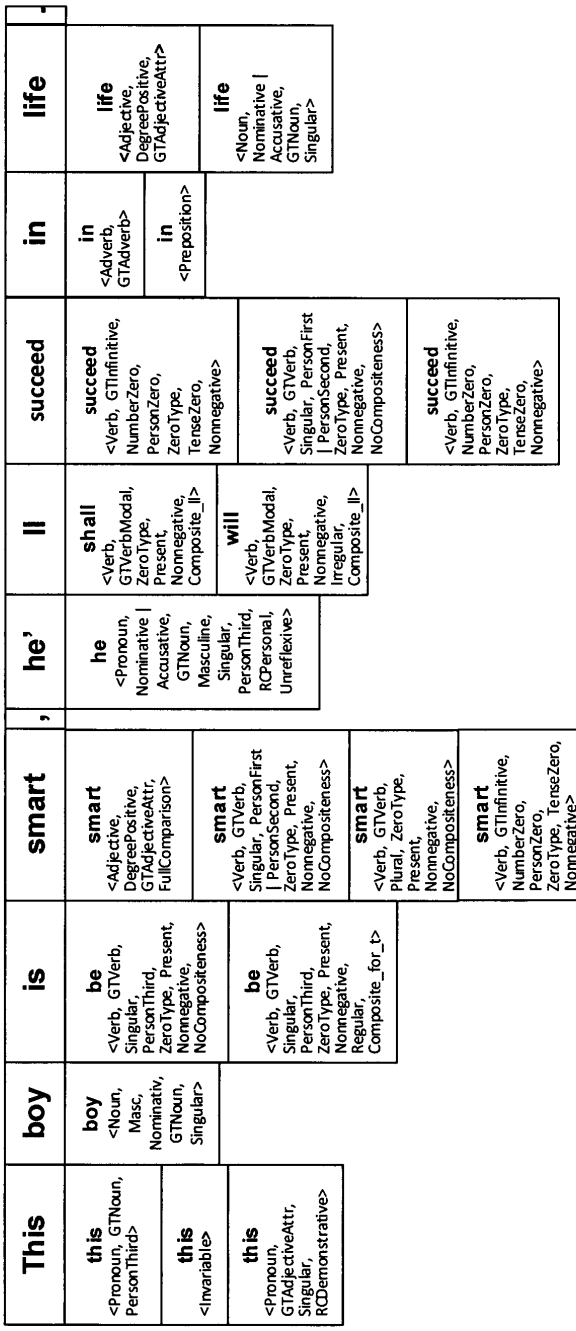
Фиг. 1

2


200
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

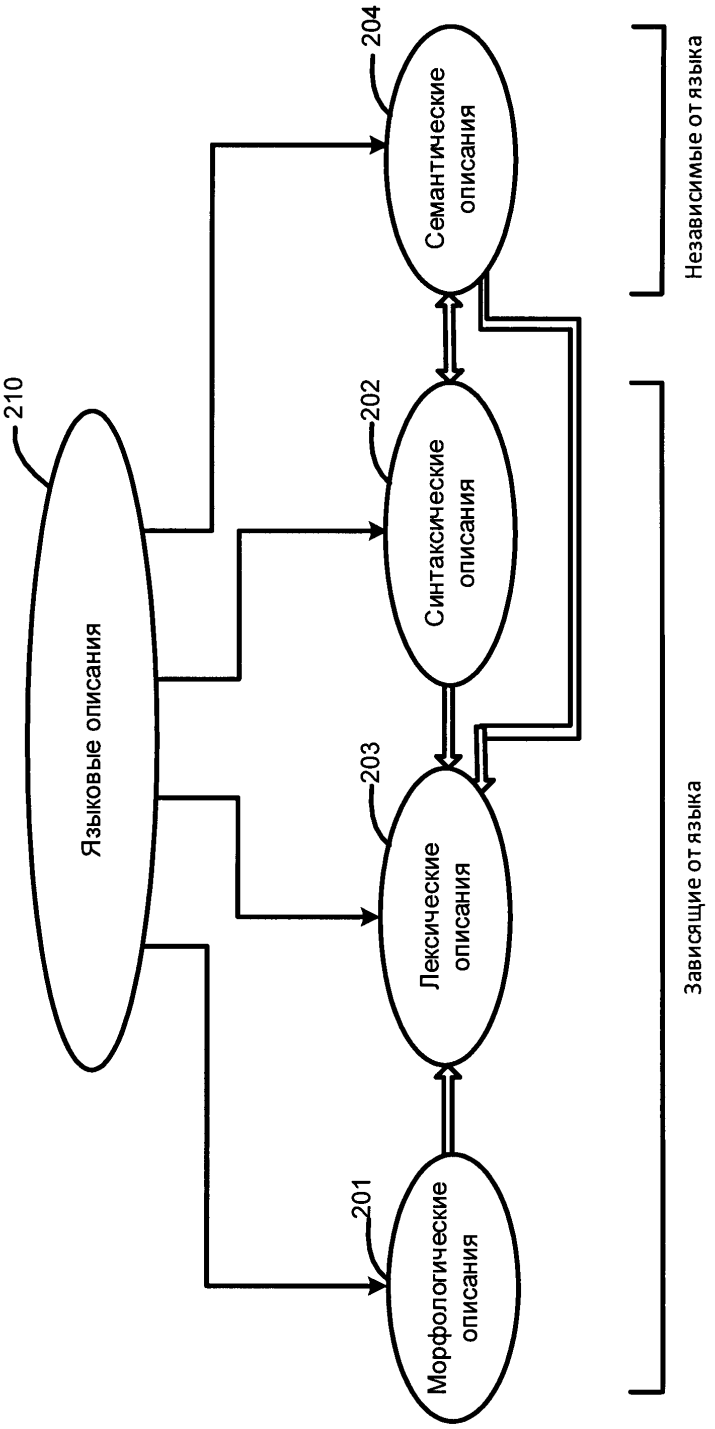
Фиг. 2

3

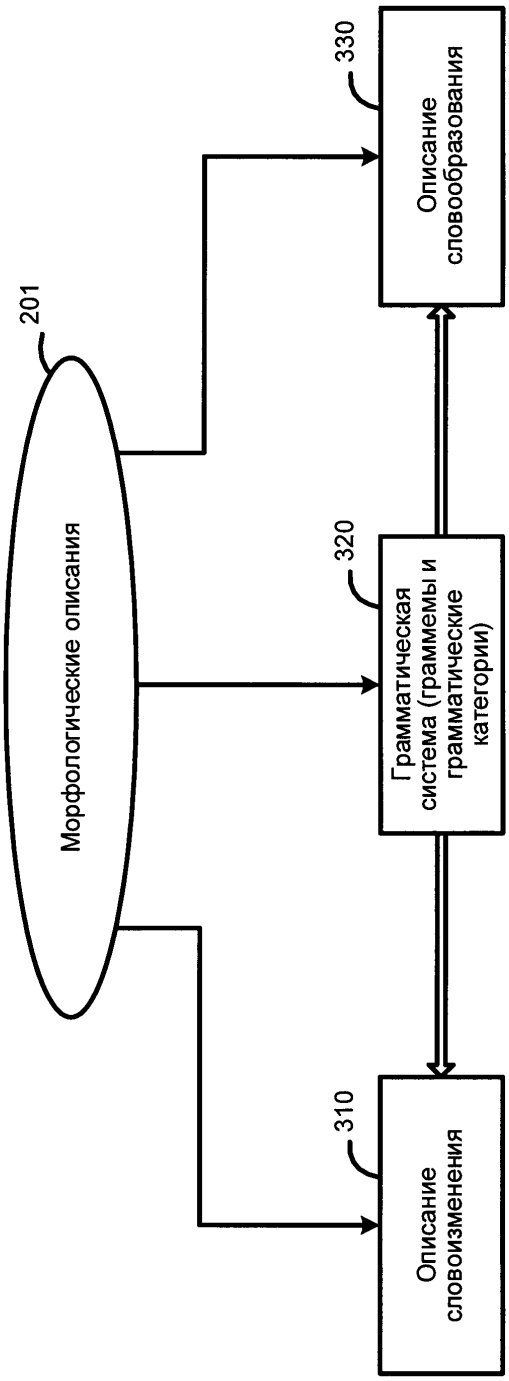

This
boy
is
smart
,
he'
ll
succeed
in
life
.
this
<Pronoun, GTNoun, PersonThird>
this
<Invariable>
this
<Pronoun, GTAdjectiveAttr, Singular, RCDemonstrative>
boy
<Noun, Masc, Nominativ, GTNoun, Singular>
be
<Verb, GTVerb, Singular, PersonThird, ZeroType, Present, Nonnegative, NoCompositeness>
be
<Verb, GTVerb, Singular, PersonThird, ZeroType, Present, Nonnegative, Regular, Composite_for_t>
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_ll>
will
<Verb, GTVerbModal, ZeroType, Present, Nonnegative, Irregular, Composite_ll>
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. 3

4

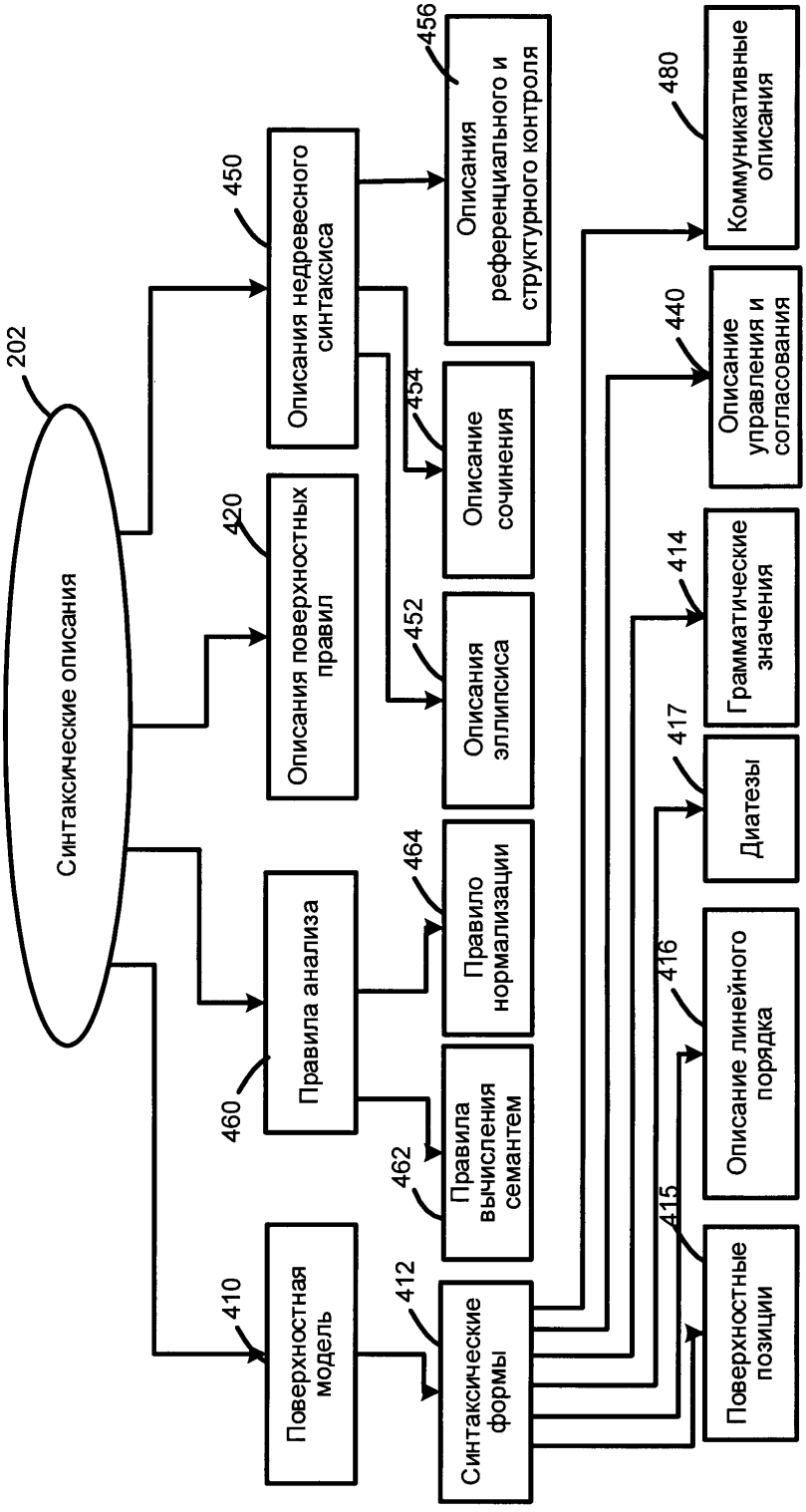

210
Языковые описания
204
Семантические описания
202
Синтаксические описания
203
Лексические описания
201
Морфологические описания
Независимые от языка
Зависящие от языка

Фиг. 4

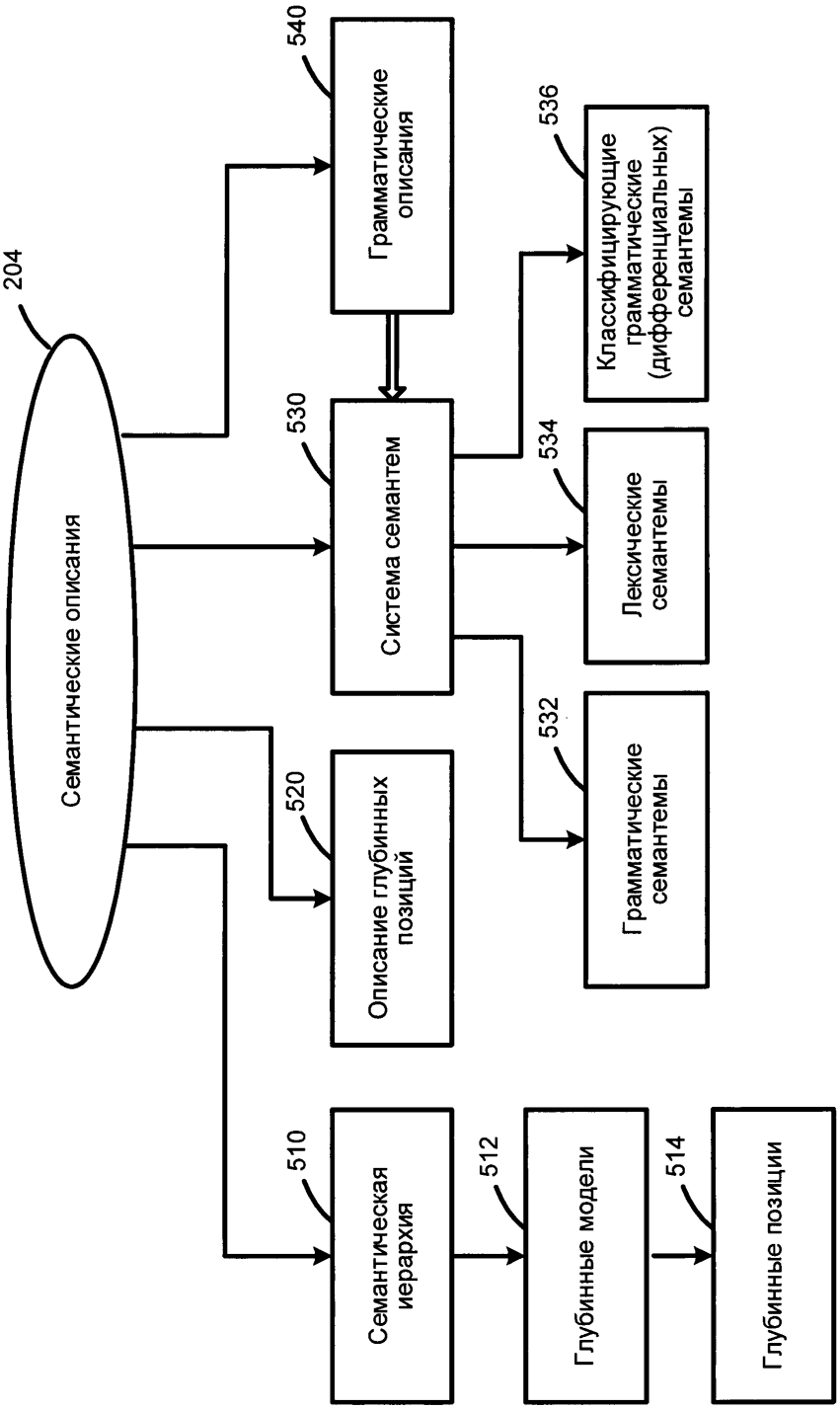

201
Морфологические описания
330
Описание словообразования
320
Грамматическая система (граммемы и грамматические категории)
310
Описание словоизменения

Фиг. 5

6

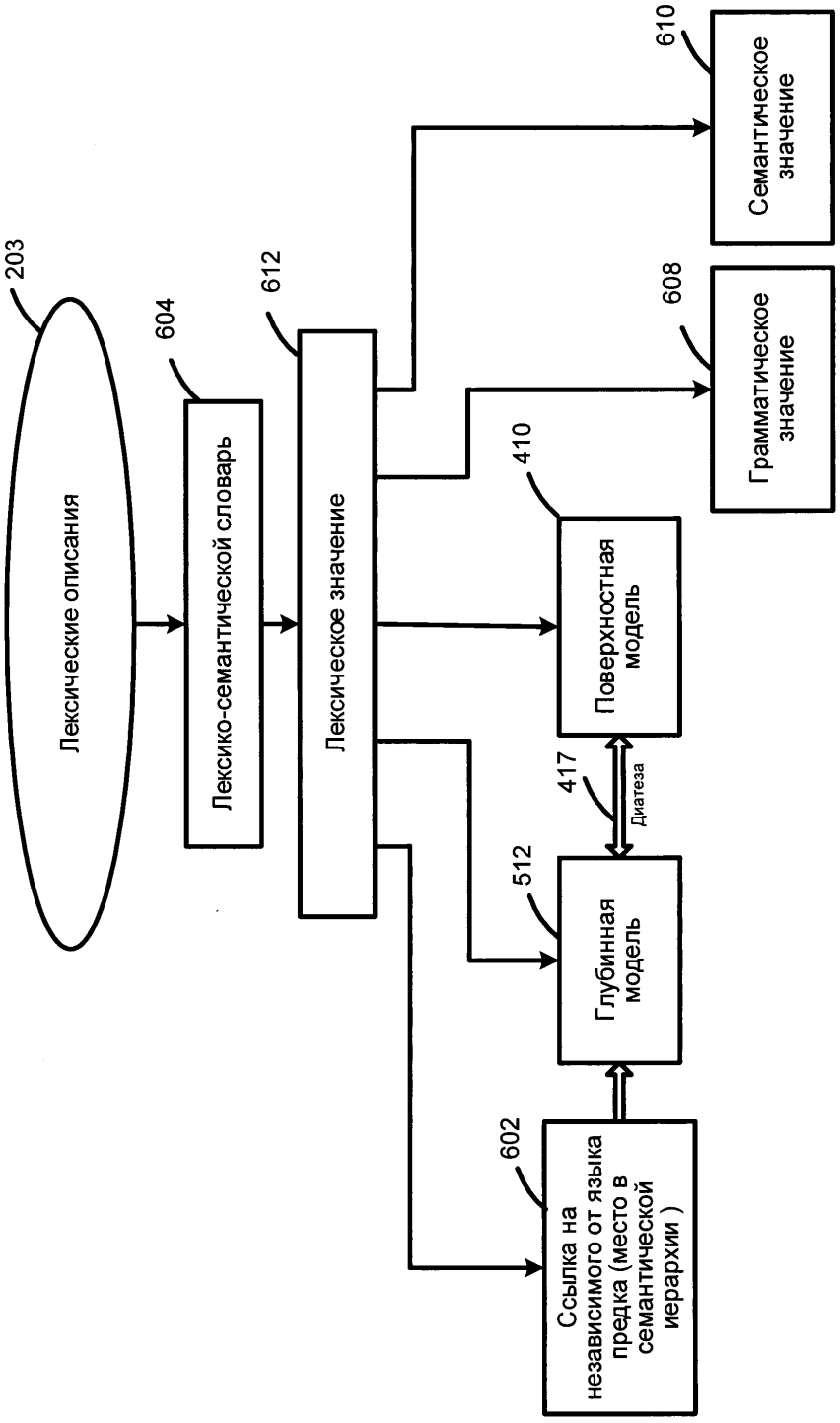

202
Синтаксические описания
410
Поверхностная модель
412
Синтаксические формы
460
Правила анализа
462
Правила вычисления семантем
464
Правило нормализации
420
Описания поверхностных правил
450
Описания недревесного синтаксиса
452
Описания эллипсиса
454
Описание сочинения
456
Описания референциального и структурного контроля
415
Поверхностные позиции
416
Описание линейного порядка
417
Диатезы
414
Грамматические значения
440
Описание управления и согласования
480
Коммуникативные описания

Фиг. 6


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

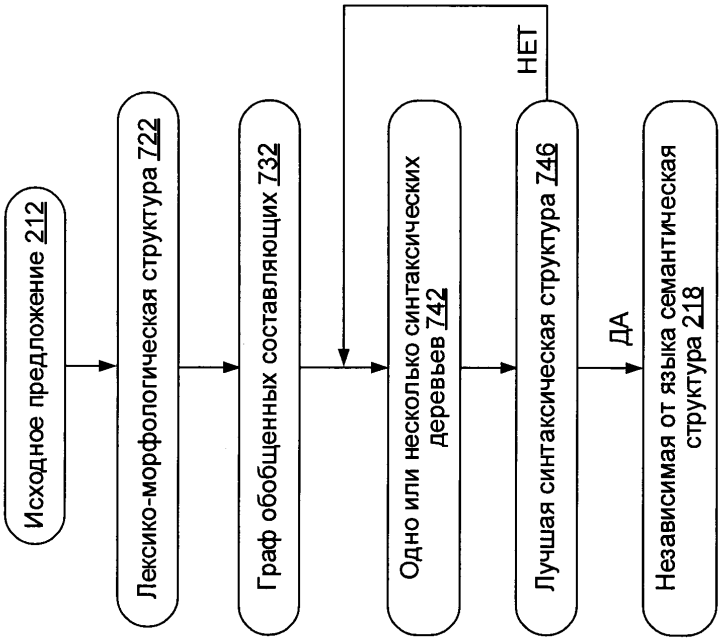
Фиг. 7

8

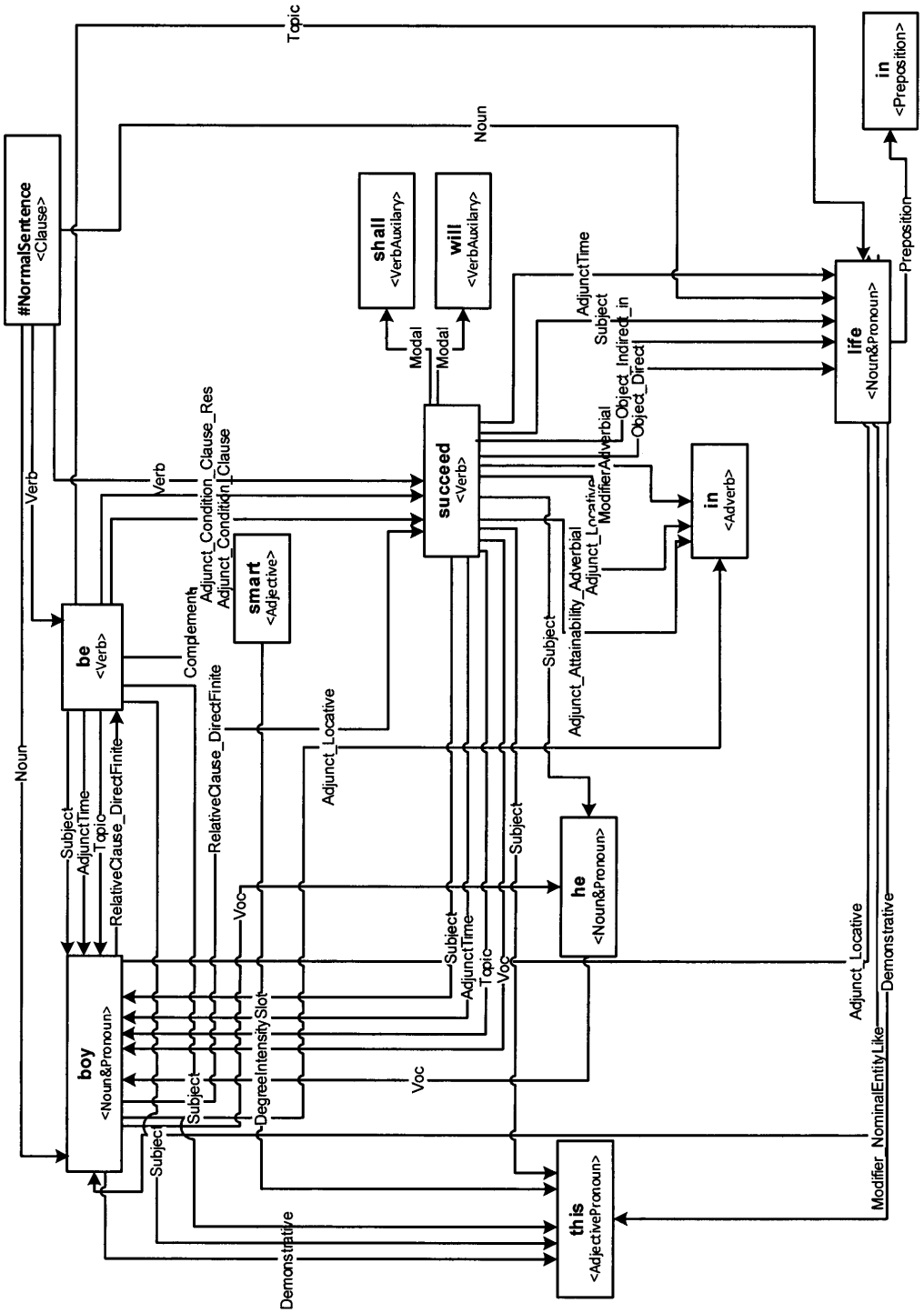

203
Лексические описания
604
Лексико-семантической словарь
612
Лексическое значение
610
Семантическое значение
608
Грамматическое значение
410
Поверхностная модель
417
Диатеза
512
Глубинная модель
602
Ссылка на независимого от языка предка (место в семантической иерархии)

Фиг. 8

9

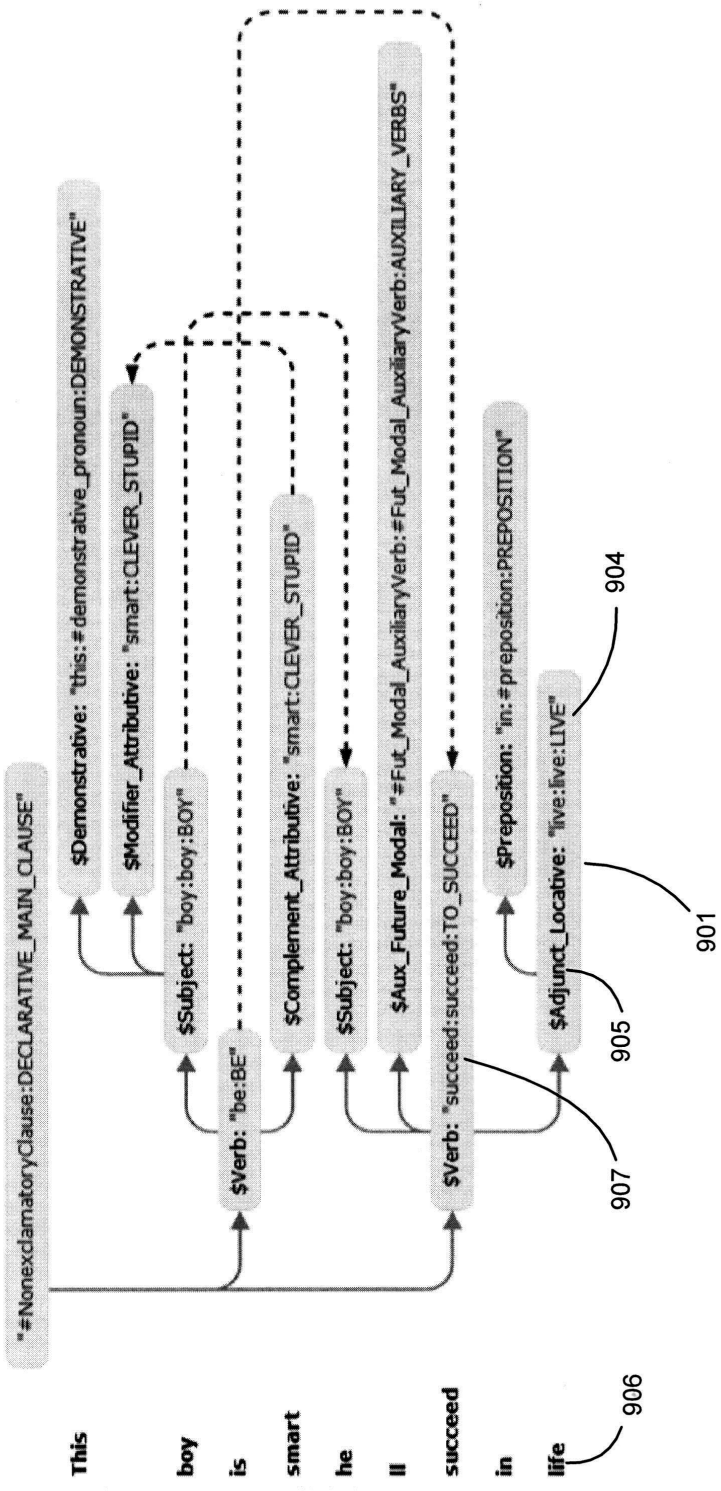

Исходное предложение 212
Лексико-морфологическая структура 722
Граф обобщенных составляющих 732
Одно или несколько синтаксических деревьев 742
Лучшая синтаксическая структура 746
НЕТ
ДА
Независимая от языка семантическая структура 218

Фиг. 9


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

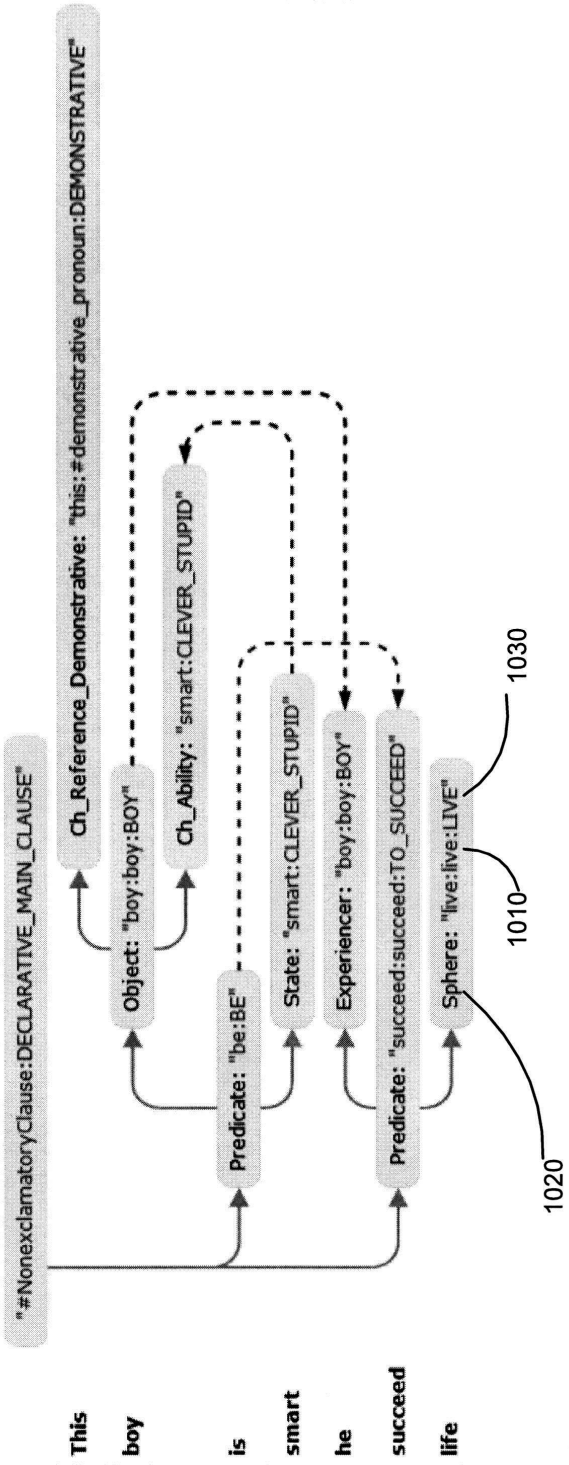
Фиг. 10

11


"#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
904
905
906
907

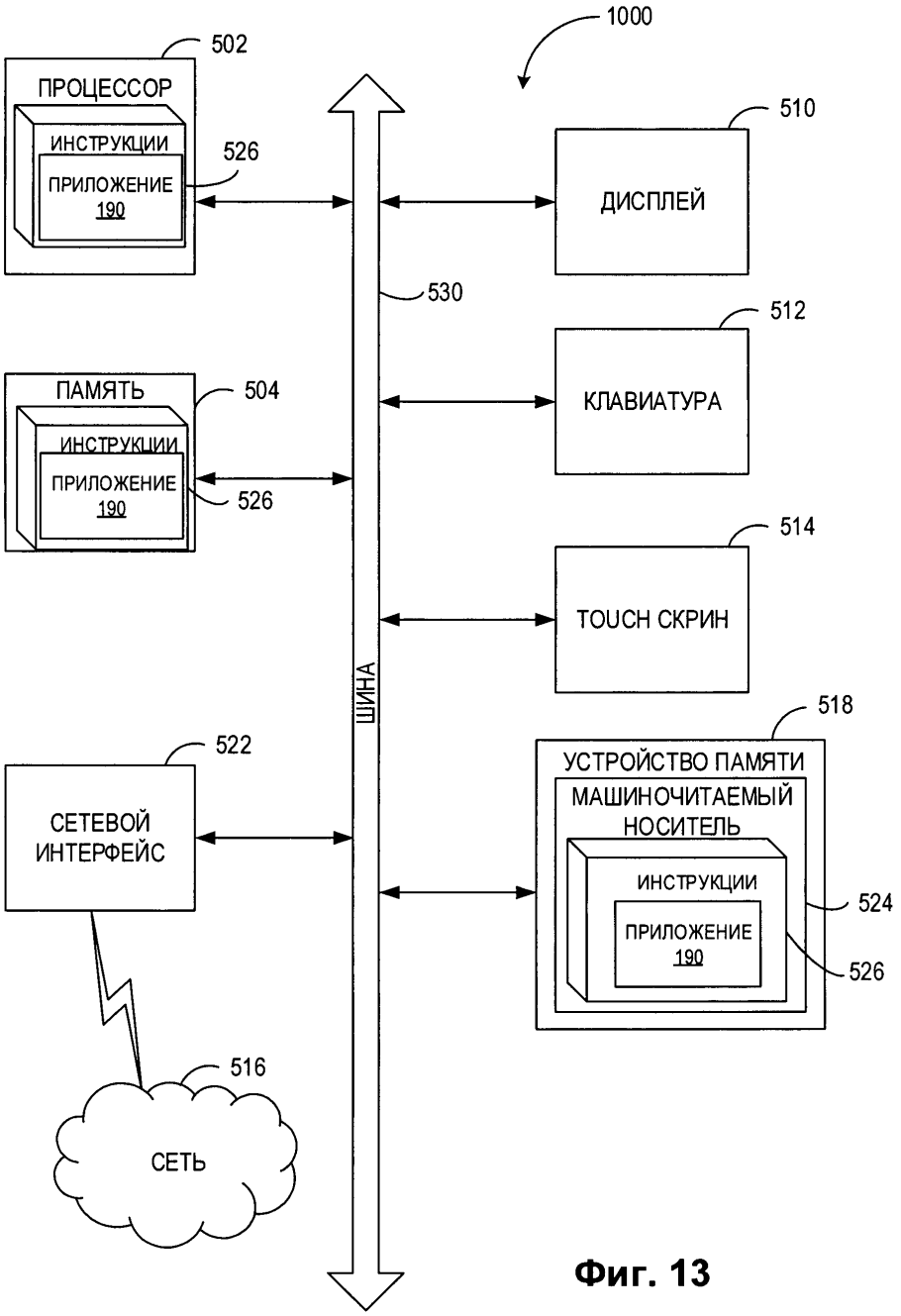
Фиг. 11

12


"#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. 12

13


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. 13