РОССИЙСКАЯ ФЕДЕРАЦИЯ

(19) **RU** (11) **2 586 577**(13) **C2**

(51) МПК
*G06F 17/21* (2006.01)

ФЕДЕРАЛЬНАЯ СЛУЖБА
ПО ИНТЕЛЛЕКТУАЛЬНОЙ СОБСТВЕННОСТИ

# (12) ОПИСАНИЕ ИЗОБРЕТЕНИЯ К ПАТЕНТУ

(21)(22) Заявка: **2014101124/08, 15.01.2014**

(24) Дата начала отсчета срока действия патента:
**15.01.2014**

Приоритет(ы):
(22) Дата подачи заявки: **15.01.2014**

(43) Дата публикации заявки: **20.07.2015** Бюл. № **20**

(45) Опубликовано: **10.06.2016** Бюл. № **16**

(56) Список документов, цитированных в отчете о поиске: **US 8195447 B2, 05.06.2012. RU 2445682 C2, 20.03.2012. US 8548795 B2, 01.10.2013. US 2012/0232883 A1, 13.09.2012. US 2013/0211816 A1, 15.08.2013.**

Адрес для переписки:
**119019, Москва, Гоголевский б-р, 11, этаж 3, "Гоулингз Интернэшнл Инк.", Соболев Александр Юрьевич**

(72) Автор(ы):
**Анисимович Константин Владимирович (RU),**
**Зуев Константин Алексеевич (RU)**

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

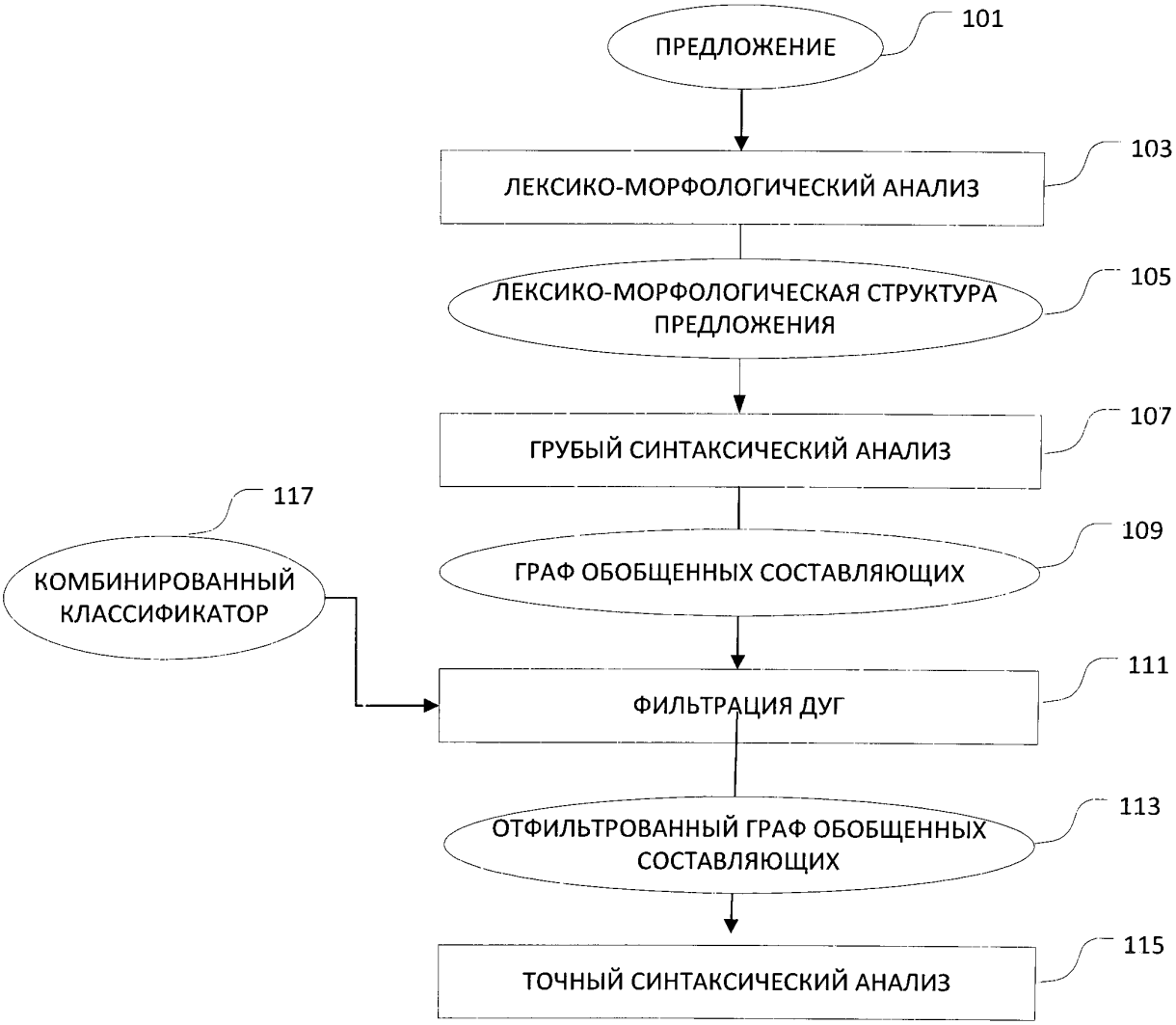
## (54) ФИЛЬТРАЦИЯ ДУГ В СИНТАКСИЧЕСКОМ ГРАФЕ

(57) Реферат:

Изобретение относится к выполнению синтаксического анализа текста. Технический результат - оценка всех возможных синтаксических комбинаций быстро и без потери истинного смысла текста. Для этого в некоторых вариантах осуществления этот способ включает выполнение грубого синтаксического анализа текста, построение графа обобщенных составляющих текста и фильтрацию дуг графа обобщенных составляющих с помощью комбинированного классификатора, который включает древесный классификатор и один или несколько линейных классификаторов. Комбинированный классификатор обучается с использованием параллельного анализа неразмеченных двуязычных текстовых корпусов. 5 н. и 18 з.п. ф-лы, 5 ил.

RU 2586577 C2

RU 2586577 C2




**Фиг. 1**

RU 2586577 C2

RU 2586577 C2

RUSSIAN FEDERATION

FEDERAL SERVICE
FOR INTELLECTUAL PROPERTY

(19) **RU** (11) **2 586 577**(13) **C2**

(51) Int. Cl.
*G06F 17/21* (2006.01)

## (12) ABSTRACT OF INVENTION

(21)(22) Application: **2014101124/08, 15.01.2014**

(24) Effective date for property rights:
**15.01.2014**

Priority:
(22) Date of filing: **15.01.2014**

(43) Application published: **20.07.2015** Bull. № **20**

(45) Date of publication: **10.06.2016** Bull. № **16**

Mail address:
**119019, Moskva, Gogolevskij b-r, 11, etazh 3, "Goulingz Interneshnl Ink.", Sobolev Aleksandr JUrevich**

(72) Inventor(s):
**Anisimovich Konstantin Vladimirovich (RU), Zuev Konstantin Alekseevich (RU)**

(73) Proprietor(s):
**Obshshestvo s ogranichennoj otvetstvennostyu "Abi InfoPoisk" (RU)**

### (54) FILTERING ARCS PARSER GRAPH

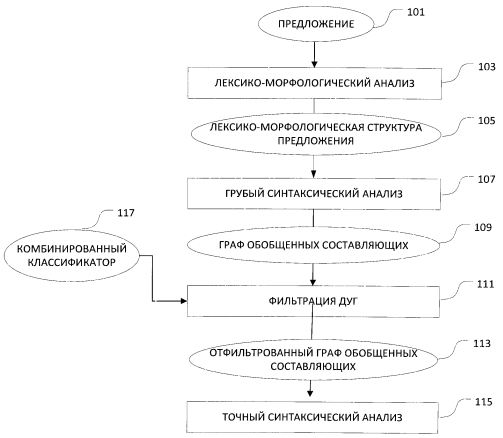
(57) Abstract:

FIELD: linguistics.

SUBSTANCE: invention relates to text parsing. For this purpose, in some versions, present method includes rough parsing of text, plotting a graph of generalised text components and filtration arcs of graph of generalised component using a combined classifier, which includes a tree classifier and one or more linear classifiers. Combined classifier is trained using parallel analysis of non-marked bilingual text corpuses.

EFFECT: technical result is evaluation of all possible syntax combinations quickly and without loss of true meaning of text.

23 cl, 5 dwg




**Фиг. 1**

RU 2586577 C2

RU 2586577 C2

## ОБЛАСТЬ ИЗОБРЕТЕНИЯ

[0001] Настоящее изобретение относится к анализу синтаксической структуры текста с помощью компьютера.

## ИЗВЕСТНЫЙ УРОВЕНЬ ТЕХНИКИ

[0002] Известно множество систем для обработки письменных текстов на естественных языках. Поскольку в структуре естественного языка присутствует неоднозначность, такие тексты непросто анализировать. Количество потенциальных синтаксических связей в предложении может быть очень большим.

[0003] Все это приводит к тому, что существует большое количество потенциальных синтаксических конструкций, которые следует рассматривать при анализе текста. Чем более продвинутой является система анализа и чем больше возможностей она учитывает, тем больше вариантов приходится анализировать. Попытка проанализировать все связи приводит к так называемому «комбинаторному взрыву». Попытка рассмотреть только наиболее вероятные комбинации может привести к потере смысла. В результате возникает проблема оптимизации процесса синтаксического анализа, которая заключается в сведении к минимуму требуемого времени при сохранении целостности результата, причем решение этой проблемы имеет критически важное значение. Настоящее изобретение предлагает способ оценки всех возможных синтаксических комбинаций быстро и без потери истинного смысла текста.

## РАСКРЫТИЕ ИЗОБРЕТЕНИЯ

[0004] В этом документе раскрываются способы, методы и системы анализа исходного предложения путем выявления предложения, определения графа обобщенных составляющих предложения, полученного на этапе грубого синтаксического анализа лексико-морфологической структуры предложения, причем граф обобщенных составляющих содержит дуги, узлы; и производится фильтрация дуг графа обобщенных составляющих с использованием комбинированного классификатора. В одном из вариантов осуществления данного изобретения комбинированный классификатор содержит древесный классификатор и по меньшей мере один линейный классификатор; причем построение синтаксической структуры исходного предложения проводится на этапе точного синтаксического анализа, использующего граф обобщенных составляющих с отфильтрованными дугами.

[0005] В одном из возможных вариантов осуществления изобретения древесный классификатор распределяет дуги по кластерам на основе заранее определенного набора признаков. В другом варианте осуществления изобретения выбранный набор признаков определяется по результатам параллельного анализа двуязычных текстовых корпусов. В другом варианте осуществления изобретения порядок признаков из предварительно заданного набора определяется согласно значению энтропии признаков. В другом варианте осуществления древесный классификатор основан на Iterative Dichotomiser 3 (итерационном дихотомическом алгоритме ID3). В некоторых вариантах осуществления веса для линейного классификатора определяются по результатам параллельного анализа неразмеченных двуязычных текстовых корпусов.

[0006] В данном документе также раскрываются способы, методы и системы для обучения классификатора, включающие выполнение параллельного анализа двуязычных текстовых корпусов, выявление двух параллельных предложений в неразмеченных двуязычных текстовых корпусах; построение двух соответствующих графов обобщенных составляющих для каждого из двух параллельных предложений; построение по меньшей мере одного синтаксического дерева на основе графов обобщенных составляющих для двух параллельных предложений, причем дуга графа обобщенных составляющих

включена в синтаксическое дерево на основании наличия соответствующей дуги у второго графа обобщенных составляющих, а также обучение комбинированного классификатора на основе построенного синтаксического дерева.

## КРАТКОЕ ОПИСАНИЕ ЧЕРТЕЖЕЙ

[0007] Дополнительные цели, характеристики и преимущества настоящего изобретения будут раскрыты в приведенном ниже описании вариантов осуществления изобретения со ссылкой на прилагаемые чертежи, в которых:

[0008] на Фиг. 1 изображена блок-схема, иллюстрирующая процесс анализа синтаксической структуры.

[0009] На Фиг. 2 изображена блок-схема, иллюстрирующая процесс грубого синтаксического анализа.

[0010] На Фиг. 3 схематически представлен граф обобщенных составляющих.

[0011] На Фиг. 4 показана блок-схема процесса фильтрации дуг.

[0012] На Фиг. 5 приведен пример компьютерной системы, которая может использоваться для реализации настоящего изобретения.

## ПОДРОБНОЕ ОПИСАНИЕ

[0013] В приведенном ниже описании для объяснения многие конкретные детали изложены для того, чтобы обеспечить полное понимание настоящего изобретения. Однако специалистам в данной области техники будет очевидно, что это изобретение может быть осуществлено практически без этих конкретных деталей. В других случаях структуры и устройства показаны только в виде блок-схем, чтобы не затруднять понимание этого изобретения.

[0014] Ссылка в этом описании на «один вариант осуществления» или «вариант осуществления» означает, что конкретный признак, структура или характеристика, описанная в связи с этим вариантом осуществления, включена по меньшей мере в один вариант осуществления настоящего изобретения. Выражение «в одном из вариантов осуществления» в различных местах описания необязательно относятся к одному и тому же варианту осуществления, также не является отдельным или альтернативным вариантом осуществления, взаимоисключающим другие варианты осуществления. Кроме того, дано описание различных признаков, которые могут использоваться в одних вариантах осуществления изобретения, и не использоваться в других. Аналогично, приведены описания различных требований, которые могут выполняться в некоторых вариантах осуществления, и не выполняться в дргуих.

[0015] На Фиг. 1 показан способ выполнения синтаксического анализа предложения в соответствии с аспектами данного изобретения. Прежде всего, система находит исходное предложения 101. Для исходного предложения 101 производится лексико-морфологический анализ 103 для построения лексико-морфологической структуры 105 предложения 101. После этого производится грубый синтаксический анализ 107 предложения 101, с использованием лексико-морфологической структуры 105. При проведении грубого синтаксического анализа 107 предложения 101 строится граф обобщенных составляющих 109 для предложения 101.

[0016] Граф обобщенных составляющих предложения является представлением всех возможных связей между словами в этом предложении. Узлы графа представляют составляющие предложения. Составляющая - это слово или группа слов, которая выступает как единое целое внутри иерархической структуры. Каждая составляющая в предложении представлена узлом на графе обобщенных составляющих. Обобщенные, например, по части речи узлы могут представлять много вариантов лексических и грамматических значений представляемых ими слов. Дуги между узлами представляют

поверхностные (синтаксические) позиции, выражающие различные типы отношений между лексическими значениями.

[0017] Граф обобщенных составляющих 109 построен из лексико-морфологической структуры 105 предложения 101. Применяются все возможные поверхностные синтаксические модели для каждого элемента лексико-морфологической структуры 105, и строятся и обобщаются все возможные составляющие. Грубый синтаксический анализатор или его эквиваленты применяется для выявления всех потенциально возможных синтаксических связей в предложении, что находит свое выражение в построении графа обобщенных составляющих 109 на основе лексико-морфологической структуры 105 с помощью поверхностных моделей, глубинных моделей, а также лексико-семантического словаря.

[0018] В одном варианте осуществления рассматриваются и обобщаются все возможные синтаксические описания и синтаксические структуры для предложения 101. В результате строится граф обобщенных составляющих 109, в котором каждая составляющая обобщена из всех возможных составляющих для каждого элемента предложения 101, а построение обобщенных составляющих выполнено для всех элементов предложения 101. Граф обобщенных составляющих 109 отражает на уровне поверхностной модели все гипотетические возможные синтаксические отношения между словами предложения 101.

[0019] При построении всех возможных составляющих каждый элемент исходного предложения 101, который не является пробелом или знаком пунктуации, рассматривается в качестве потенциального ядра составляющей. Построение графа обобщенных составляющих 109 начинается с построения тех составляющих, которые имеют только словоформу, являющуюся ядром, а затем они расширяются для построения составляющих на следующем уровне путем включения соседних составляющих. Для каждой пары «лексическое значение грамматическое значение», инициализируется ее поверхностная модель, прикрепляются другие составляющие в поверхностных позициях синтформ ее поверхностной модели к правым и левым соседним составляющим. Если соответствующая синтформа обнаруживается в поверхностной модели соответствующего лексического значения, то выбранное лексическое значение может быть ядром новой составляющей.

[0020] Граф обобщенных составляющих 109 первоначально строится в виде дерева, от листьев к корню (снизу вверх). Построение дополнительных составляющих производится снизу вверх путем присоединения дочерних составляющим к родительским составляющих посредством заполнения поверхностных позиций родительских составляющих, чтобы покрыть все начальные лексические единицы предложения 101.

[0021] Корень дерева является главной частью, представляющей специальную составляющую, которая соответствует различным типам максимальных единиц анализа текста (полные предложения, перечисления, названия и т.д.). Ядром главной части является, как правило, предикат (сказуемое). В ходе этого процесса дерево на самом деле становится графом, так как составляющие более низкого уровня (листья) могут быть включены в различные составляющие верхнего уровня (корень).

[0022] Некоторые из составляющих, которые строятся для того же самого элемента лексико-морфологической структуры, могут быть обобщены для получения обобщенных составляющих. Составляющие обобщаются, среди прочих, на основе лексических значений, грамматических значений, например, на основе частей речи, своих связей. Составляющие обобщаются границами (связями), так как существует множество различных синтаксических связей в предложении, и одно и то же слово может быть

включено в различные составляющие. В качестве результата грубого синтаксического анализа 107 строится граф 109 обобщенных составляющих, который представляет все предложение 101 целиком.

[0023] На Фиг. 2 приведена более подробная иллюстрация грубого синтаксического анализа 107 в соответствии с одним или несколькими вариантами осуществления изобретения. Грубый синтаксический анализ 107 в числе основных операций включает предварительный сбор 201 составляющих, построение обобщенных составляющих 203, фильтрацию 205, построение моделей обобщенных составляющих 207, построение графа обобщенных составляющих 209, обработку согласования 211, восстановление эллипсиса 213.

[0024] Предварительный сбор 201 составляющих на этапе грубого синтаксического анализа 107 выполняется на основе лексико-морфологической структуры 105 анализируемого предложения, включая определенные группы слов, слова в скобках, кавычках и т.д. них составляющих в рамках выбранной модели. Предварительный сбор 201 выполняется в начале грубого синтаксического анализа 107 до построения обобщенных составляющих 203 и построения обобщенных составляющих 207, чтобы охватить все связи в предложении.

[0025] Для построения обобщенных составляющих 203 обычно требуется, чтобы все возможные пары лексических значений и грамматических значений были найдены или назначены каждой составляющей, а поверхностные позиции дочерних составляющих были присвоены каждой из этих составляющих. Лексические единицы предложения 101 могут сформировать ядро составляющих на нижних уровнях. Каждая составляющая может быть прикреплена к составляющей более высокого уровня, если поверхностные позиции составляющей более высокого уровня могут быть заполнены. Таким образом, составляющие дополнительно расширяются, чтобы включить соседние составляющие, построенные в ходе более раннего процесса построения составляющих, пока не будут построены все возможные составляющие, покрывающие все предложение.

[0026] При грубом синтаксическом анализе 107 число различных составляющих, которые могут быть построены, и число синтаксических отношений между ними достаточно велико, поэтому некоторые из поверхностных моделей составляющих выбираются, чтобы быть отсортированными в процессе фильтрации 205 до и после построения составляющих или в ходе такого построения для того, чтобы уменьшить количество различных рассматриваемых составляющих.

[0027] Фильтрация 205 при грубом синтаксическом анализе 107 включает фильтрацию набора синтформ 215, выполненную до построения обобщенных составляющих 203 или во время их построения. Синтформы 215 и поверхностные позиции фильтруются априори, а составляющие фильтруются после того как они построены. Процесс фильтрации 205 позволяет исключить ряд синтформ 215, включая следующие, но не ограничиваясь ими: те синтформы, которые не соответствуют грамматическим значениям этой составляющей, те синтформы, в которых ни одна из позиций ядра не может быть заполнена, синтформы со специальными позициями, которые описывают грамматические движения. Специальная позиция, такая как образование относительного придаточного предложения, и вопрос, предполагающая специальную лексему (относительное или вопросительное местоимение), отфильтровывается, если специальная лексема отсутствует в предложении. Процесс фильтрации 205 позволяет существенно уменьшить число рассматриваемых вариантов разбора. Однако существуют и маловероятные варианты значений, поверхностных моделей и синтформ, исключение которых из последующего рассмотрения может привести к потере маловероятного,

но, тем не менее, возможного смысла.

[0028] Обычно те синтаксические формы (синтформы 215), которые не имеют заполнителя по меньшей мере для одной поверхностной позиции, могут быть отфильтрованы и отброшены. Кроме того, те лексические значения, которые не имеют синтформ 215 с заполненными поверхностными позициями, отвергаются и фильтруются. Грубый синтаксический анализ 107 невозможно выполнить, если нет синтформ с заполненной поверхностной позицией как таковой, производится фильтрация 205.

[0029] После того как построены все возможные составляющие, выполняется процедура обобщения для построения обобщенных составляющих 207. Все возможные омонимы и все возможные значения элементов исходного предложения, которые могут быть представлены одной и той же частью речи, собираются и обобщаются, и все возможные построенные таким образом составляющие группируются в обобщенные составляющие 217.

[0030] Обобщенная составляющая 217 описывает все составляющие со всеми возможными связями в исходном предложении, которое имеет некоторую словоформу в качестве ядра, включая различные лексические значения этой словоформы. Поскольку составляющие обобщаются, строится общая составляющая для всех лексических значений, соответствующих данной словоформе, в том числе омонимы, и их синтаксические формы могут рассматриваться одновременно.

[0031] Далее выполняется построение 207 моделей обобщенных составляющих 219, включающих обобщенные модели всех обобщенных лексем. Модели обобщенных составляющих лексем содержат обобщенные глубинные модели и обобщенные поверхностные модели. Обобщенная глубинная модель лексем включает перечень всех глубинных позиций, а также описания всех требований к заполнителям глубинных позиций. Обобщенная поверхностная модель содержит информацию о синтформах 215, в которых может встречаться данная лексема, о поверхностных позициях, о диатеах (соответствиях между поверхностными позициями и глубинными позициями), а также описание линейного порядка.

[0032] Синтформы 215 и поверхностные позиции, которые являются важными для этой лексемы, выбираются с помощью битовой маски. Кроме того, строятся модели обобщенных составляющих, потому что составляющая обобщается не только по лексическим значениям и синтаксическим формам своего ядра, но и по тем фрагментам, которые она заполняет. Использование моделей обобщенных составляющих уменьшает количество нерелевантных связей и помогает оптимизировать процесс извлечения синтаксического дерева таким образом, чтобы учесть все возможные границы.

[0033] Как показано на Фиг. 2, информация от синтформ 215 синтаксического описания, а также семантические описания используются для построения моделей обобщенных составляющих 219. Например, дочерние составляющие присоединяются к каждому лексическому значению лексической единицы, и грубый синтаксический анализ 107 может быть необходим для определения того, может ли составляющая более низкого уровня быть заполнителем соответствующей поверхностной позиции в составляющей верхнего уровня. Такой сравнительный анализ позволяет отсечь на ранней стадии неправильные синтаксические связи.

[0034] Далее проводится построение графа обобщенных составляющих 209. Граф обобщенных составляющих 109, который описывает все возможные синтаксические структуры всего предложения, строится в результате сбора обобщенных составляющих 217. Построение графа обобщенных составляющих 209 организуется путем формирования и обработки очереди запросов для прикрепления одной составляющей

к другой составляющей. В общем случае пары составляющих, представляющих контактные группы слов в предложении, могут быть включены в очередь запросов.

[0035] Фиг. 3 представляет собой пример графа обобщенных составляющих 109 для предложения «This boy is smart, he'll do well in life.». Составляющие, показанные прямоугольниками, представляют лексему, являющуюся ядром соответствующей составляющей. Морфологическая парадигма (как правило, это часть речи) ядра составляющей выражена граммемами частей речи и изображается в угловых скобках под лексемой. Морфологическая парадигма, как часть описания словоизменений морфологического описания, содержит всю информацию об изменении формы слова одной или нескольких частей речи. Например, поскольку словоформа «do» может принадлежать двум частям речи: существительному <Noun> и глаголу <Verb> (что представлено обобщенной морфологической парадигмой <Noun&Pronoun>), на графе 109 показаны две составляющие для слова «do».

[0036] Связи в графе 109 представляют собой заполненные поверхностные позиции составляющих. Названия позиций отображаются на стрелках графа. Любая составляющая сформирована ядром лексемы, которая может иметь исходящие именованные стрелки, обозначающие поверхностные позиции, заполненные дочерними составляющими. Входящая стрелка обозначает заполнение данной составляющей поверхностной позиции в другой составляющей. Граф 109 настолько сложен и имеет так много стрелок (ветвей) в связи с тем, что он отображает все возможные связи, которые могут быть установлены между словами предложения «This boy is smart, he'll do well in life.». В графе 109, однако, имеется множество связей, которые будут отброшены. Значение грубой оценки сохраняется при каждой стрелке, обозначающей заполненную поверхностную позицию. Только поверхностные позиции и связи с высоким значением рейтинговых оценок будут выбраны в первую очередь на следующем этапе синтаксического анализа.

[0037] Зачастую несколько стрелок могут соединять одни и те же пары составляющих. Это означает, что существует несколько подходящих поверхностных моделей для этой пары составляющих, и несколько поверхностных позиций родительской составляющей могут быть альтернативно заполнены этой дочерней составляющей. Так, три поверхностные позиции с именами Idiomatic_Adverbial 301, Modifier_Adverbial 303 и AdjunctTime 305 родительской составляющей «do<Verb»> 307 могут быть независимо заполнены дочерней составляющей «well<Adverb>» 309 в соответствии с поверхностной моделью составляющей «do<Verb>». Таким образом, грубо говоря, "do<Verb>" 307 и "well<Adverb>" 309 формируют новую составляющую с ядром «do<Verb>», которая присоединена к другой родительской составляющей, например, к составляющей #NormalSentence<Clause>311 в поверхностной позиции Verb 313, и к составляющей «child<Noun&Pronoun>» 315 в поверхностной позиции.

RelativClause_DirectFinite 317. Помеченный элемент #NormalSentence<Clause>311, является «корнем», он соответствует всему предложению.

[0038] Описание процесса грубого синтаксического анализа и лексико-морфологического анализа приведено, например, в патенте США №8,078,420, ссылка на который дана в этом документе.

[0039] Возвратимся к Фиг. 1; на этапе грубого синтаксического анализа 107 предложения 101 проводится построение графа обобщенных составляющих 109. Затем итоговый граф 109 подвергают фильтрации дуг 111 для того, чтобы существенно уменьшить количество дуг в графе 109. В некоторых вариантах осуществления отфильтрованный граф обобщенных составляющих 113 затем используется для

выполнения точного синтаксического анализа 115. Отфильтрованный граф 113 содержит существенно меньшее число дуг по сравнению с графом 109 до фильтрации. Снижение количества дуг в графе 109 приводит к увеличению производительности и повышению качества точного синтаксического анализа 113.

[0040] Фильтрация дуг 111 графа обобщенных составляющих 109 выполняется комбинированным классификатором 117.

[0041] Классификатор представляет собой алгоритм классификации набора элементов. Классификацией называется процесс разделения множества элементов на выбранные каким-либо образом классы.

[0042] В некоторых вариантах осуществления комбинированный классификатор 117 представляет собой комбинацию линейного классификатора и древесного классификатора. Комбинированный классификатор 117 представляет собой дерево решений с линейными классификаторами 405 в его узлах. Комбинированный классификатор 117 фильтрует дуги с использованием символьных и числовых признаков.

[0043] Символьные признаки представляют собой идентификаторы, соответствующие элементам описания. Например, в некоторых вариантах осуществления следующие признаки и их сочетания, среди прочих, используются в качестве символьных признаков:

- поверхностная позиция, которая является способом синтаксического выражения глубинной (семантической) позиции. Например, поверхностной позицией может быть указание на член предложения, играющий указанную роль в предложении: субъект, объект, предикат и т.д.;
- длина связи, которая представляет собой число шагов, которые можно использовать, чтобы пройти по синтаксическому дереву от одного слова к другому слову с учетом знаков препинания;
- синтпарадигма 1,2, которая представляет собой набор синтаксических форм, описывающих поверхностную реализацию глубинной модели для данного ядра. Индекс обозначает слово в паре, которая образует связь.
- шаблон эллипсиса (эллиптическая лексема);
- синтаксическая парадигма 1, 2 $_{-1/+1/-2/+2}$.

[0044] В некоторых вариантах осуществления числовые признаки могут включать различные априорные и статистические оценки, такие как:

- оценка по триграммам 1;
- оценка по триграммам 2;
- статистическая оценка заполнения поверхностных позиций;
- оценка1 эллипсиса;
- оценка2 эллипсиса;
- оценка пунктуации;
- оценка управления.

[0045] Как показано на Фиг. 4, в некоторых вариантах осуществления данного изобретения комбинированный классификатор 117 получает набор дуг 401 графа обобщенных составляющих 109, древесный классификатор 403 разделяет множество этих дуг на кластеры, затем линейный классификатор 405 принимает решение о разделении дуг на «плохой» и «хороший» кластеры, отбрасывает дуги из «плохих» класеров, после чего формируется отфильтрованный набор дуг графа обобщенных составляющих 407.

[0046] Когда запускается процесс фильтрации, древесный классификатор 403 изначально делит дуги на кластеры, используя символьные признаки 409, выбранные в качестве основных признаков во время обучения классификатора 411. Число шагов

кластеризации определяется количеством символьных признаков 409, выбранных во время обучения классификатора 411, и количеством данных, которое соответствуют им.

[0047] Древесный классификатор 403 создает дерево решений, которое классифицирует дуги графа обобщенных составляющих 401. Каждый узел дерева соответствует одному из признаков множества символьных признаков 409.

[0048] В некоторых вариантах реализации данного изобретения порядок, в котором символьные признаки 409 применяются к набору дуг, определяется энтропией (или приростом информации) соответствующих признаков. В ходе каждой итерации классификатор вычисляет энтропии H(S) для каждого неиспользованного признака в множестве символьных признаков 409, где S является множеством дуг графа обобщенных составляющих 401. H(S) следует понимать как меру неопределенности или непредсказуемости информации, которую получают от множества S на основании каждого признака, и выбирают признак с наименьшей энтропией Hmin(S). Эта мера соответствует максимальному приращению информации для этого признака, то есть разности энтропии до и после разделения множества S в соответствии с этим признаком. Иными словами, прирост информации показывает, какая неопределенность в S удаляется после разбиения S в соответствии с этим признаком. Выбирается признак с наименьшей энтропией (или с максимальным приростом информации), и множество дуг графа обобщенных составляющих 401 разделяется на два подмножества так, чтобы элементы одного подмножества удовлетворяли признаку с Hmin(S), а элементы другого множества - нет.

[0049] Применение древесного классификатора 403 продолжается рекурсивно на каждом подмножестве, рассматривая только те элементы, которые не были выбраны ранее. Рекурсия подмножества может остановиться в одном из следующих случаев:

- все элементы подмножества принадлежат одному из кластеров (+ или -);
- признаки закончились, а элементы, относящиеся к разным классам, все еще содержатся в подмножестве. В этом случае узел обращается в лист и маркируется именем кластера, к которому относится большая часть элементов;
- в подмножестве не осталось элементов, удовлетворяющих выбранному на данном шаге признаку. В этом случае создается лист и маркируется именем кластера, к которому относится наибольшее число элементов родительского узла.

[0050] Способ, используемый древесным классификатором, включает следующие основные шаги:

- вычисление энтропии для каждого признака из множества данных S;
- разбиение множества S на подмножества на основе признака с минимальной энтропией (или, другими словами, с максимальным приростом информации);
- создание узлов дерева решений, которые содержат каждый признак;
- рекурсию по подмножествам с использованием остающихся признаков.

[0051] Может оказаться, что в этом процессе выделяется избыточное количество признаков. Для того чтобы избежать этого, устанавливают ограничение размера дерева. Более эффективным решением является создание меньших деревьев, нежели больших.

При этом предпочтительно создавать маленькие, но не минимально возможные деревья.

[0052] Энтропия H(S) является мерой неопределенности во множестве S.

$H(S)=-\Sigma_{x\in X}p(x)\log_2 p(x)$, где S является текущим множеством, для которого вычисляется энтропия на каждой итерации метода, x - это выбор классов в множестве S, а p(x) представляет собой отношение между числом элементов в классе x и числом элементов

в множестве S.

[0053] Если H(S)=0, то множество S идеально классифицировано в том смысле, что все элементы S принадлежат к одному классу.

[0054] Прирост информации IG(i) является мерой снижения неопределенности в множестве S после разделения его в соответствии с признаком i.

[0055]=IG(i)=H(S)-$\Sigma_{t\in T}$p(t)H(t), где H(S) - это энтропия выбора S, T является подмножеством, полученным путем разбиения S по признаку i, S=$U_{t\in T}$t; p(t) представляет собой отношение числа элементов в t к числу элементов в S, а H(t) является энтропией подмножества t.

[0056] В некоторых вариантах осуществления данного изобретения признаки в древесном классификаторе 403 упорядочиваются, например, таким образом:

- поверхностная позиция;
- длина связи;
- синтпарадигма 1,2 (индекс соответствует слову в связи)
- шаблон эллипсиса 1,2;
- синтпарадигма 1,2;
- синтпарадигма +1, -1;
- синтпарадигма +2, -2.

[0057] В некоторых вариантах осуществления данного изобретения древесный классификатор 403 основан на алгоритме Iterative Dichotomiser 3 (итерационном дихотомическом алгоритме ID3), алгоритме CART или алгоритме C4.5.

[0058] В некоторых вариантах осуществления изобретения комбинированный классификатор 117 включает один линейный классификатор 405. В других вариантах осуществления комбинированный классификатор 117 содержит два или более линейных классификатора 405.

[0059] Линейный классификатор - это классификатор, который принимает классифицирующее решение на основе линейной комбинации характеристик классифицируемых элементов. Как правило, характеристики элемента представлены в виде вектора. Если вектор характеристики $\vec{x}$ подается на вход классификатора, то выходная оценка имеет вид $y = f(\vec{w},\vec{x}) = f(\sum_j w_j x_j)$, где $\vec{w}$ - это вектор весов, а f является функцией, которая преобразует скалярное произведение векторов в требуемое выходное значение. Вектор весов $\vec{w}$ или набор весов вычисляются в ходе обучения классификатора с использованием промаркированных тренировочных примеров. Обычно f является простой функцией, которая относит все значения, превышающие некоторое пороговое значение, к первому классу, а все остальное - ко второму классу. Более сложные функции могут, например, представлять вероятность того, что элемент принадлежит к определенному классу. Для задач с распределением на два класса можно интерпретировать работу линейного классификатора как разделение многомерного пространства входных данных с помощью гиперплоскости. Все точки с одной стороны от плоскости классифицируются со значением «да», а все другие классифицируются со значением «нет».

[0060] В некоторых вариантах осуществления линейный классификатор 405 принимает в качестве входного множество дуг в графе обобщенных составляющих 401 предложения 101, которые предварительно объединены в кластеры с использованием древесного классификатора 403. Линейный классификатор 405 принимает для каждого кластера решение о том, является ли кластер дуг «хорошим» или «плохим». Линейный классификатор 405 принимает это решение о каждом кластере путем вычисления

выходной оценки для кластера на основе индивидуальных характеристик этого анализируемого кластера и вектора весов 413 классификатора, который предоставляется классификатору 117 и линейному классификатору 405 путем обучения классификатора 411.

[0061] После получения решений для всех кластеров линейный классификатор 405 комбинирует решения, принятые для каждого кластера, а затем применяет формулу Байеса к решениям. Сравниваются вероятности того, что дуги попадут в кластеры «хороших» и «плохих» дуг. Эти вероятности вычисляются, исходя из предположения о нормальном распределении проекций дуг в пространстве числовых признаков на нормаль к разделяющей плоскости. Проекция на нормаль к поверхности вычисляется как

$$x = X*W/|W|,$$

где X представляет собой вектор числовых признаков связей, а

W является вектором весов в классификаторе.

[0062] Вероятность того, что связи попадают в кластеры «хороших» и «плохих» связей, можно вычислить по следующей формуле:

$$Pgood= p(x)good/( p(x)good +p(x)bad),$$

где p(x) - это плотность вероятности нормального распределения.

[0063] Для каждого сочетания символьных признаков дуг мы знаем процент «хороших» дуг. Таким образом, полная вероятность того, что дуга является «хорошей», рассчитывается как

$$Pgood = Pgood(symb)* Pgood(x).$$

[0064] Символьные признаки 409 для древесного классификатора 403 и веса 413 для линейного классификатора 405 определяются в процессе обучения классификатора 411. В некоторых вариантах осуществления данного изобретения обучение классификатора осуществляется путем параллельного анализа 415 двуязычных текстовых корпусов.

[0065] Двуязычный текстовый корпус является неразмеченным текстовым корпусом на двух языках, например, на русском и английском языках, в котором каждое предложение на одном языке имеет соответствующее предложение на втором языке, причем одно предложением является точным переводом другого предложения.

[0066] На этапе параллельного анализа 415 для каждого предложения на двух языках сначала проводится разбор, а именно строятся графы обобщенных составляющих, и путем сравнения из двух графов вычленяются деревья, связи в которых наилучшим образом совпадают с точки зрения семантических классов. Например, если в одном графе есть связь «показать файл», в котором лексема «показать» принадлежит семантическому классу «TO SHOW», и лексема «файл» принадлежит классу «FILE», в то время как в другом графе имеется эквивалентная связь «show file», для которой «show» относится к классу «TO SHOW», а «file» включен в класс «FILE», то эта связь считается «хорошей». Например, если для «show file» из второго графа лексема «show» является существительным и принадлежит к классу «PERPORMANCE», а лексема «file» представляет собой глагол из класса «TO FILE», то эта связь является «плохой». Поскольку используемый при анализе корпус является неразмеченным, т.е., переводчик не указал «хорошие» или «плохие» дуги вручную, то в общем случае можно построить несколько таких древесных структур для каждой пары графов, соответствующих паре предложений. В результате параллельного анализа 411 на основе пар графов обобщенных составляющих строится по меньшей мере одна древесная структура, содержащая только «хорошие» дуги.

[0067] На основании параллельного анализа 415 двуязычных текстовых корпусов проводится обучение классификатора 411 с тем, чтобы научить классификатор 117 определять, является ли дуга из набора дуг графа обобщенных составляющих 401 «хорошей» или «плохой». Это обучение основано на принципе сравнения графа обобщенных составляющих с по меньшей мере одной древесной структурой, построенной с использованием параллельного анализа 415. Входной граф имеет все возможные синтаксические дуги. Затем все дуги в этом графе, которые также присутствуют по меньшей метре в одном синтаксическом дереве, построенном при параллельном анализе 415, помечаются как «хорошие», а дуги, которые отсутствуют в синтаксических деревьях, помечаются как «плохие». Выходной граф представляет собой граф с помеченными дугами. Во время обучения классификатор не только учится отличать «хорошие» дуги от «плохих», он также формирует набор символьных признаков 409, которые будут использоваться древесным классификатором 403.

[0068] Линейные классификаторы 405 в узлах синтаксического дерева комбинированного классификатора 117 могут обучаться с помощью различных методов. В некоторых вариантах осуществления эти методы включают способ минимизации функции ошибки или метод опорных векторов (метод SVM).

Обработка грамматической омонимии

[0069] В некоторых вариантах осуществления данного изобретения комбинированный классификатор 117 используется для устранения затруднений, связанных с омонимией. Если в одном из узлов графа обобщенных составляющих 109 имеется несколько грамматических омонимов, то их вероятности рассчитываются на основе пропорциональной частоты в N-граммах, что в данном случае следует понимать как последовательность слов или контекст:

$p_1=\exp(Q_1)/[\exp(Q_1)+\exp(Q_2)]$

$p_2=\exp(Q_2)/[\exp(Q_1)+\exp(Q_2)]$

[0070] Аналогично поступают с альтернативными шаблонами эллипсиса, используя для вычисления относительных частот оценки шаблонов.

[0071] Выбор пути в синтаксическом дереве осложняется грамматической омонимией и неоднозначностью шаблонов эллипсиса. Поэтому окончательное решение принимается следующим образом:

[0072] Для каждого пути в дереве вычисляют произведение вероятностей узлов в этом пути, при этом получают полную вероятность пути.

[0073] Для каждого пути в дереве рассчитывается вероятность того, что путь приведет в «хороший» кластер или в «плохой» кластер, а затем рассчитывается полная вероятность того, что этот путь приведет в «хороший» или в «плохой» кластер.

[0074] $P_{good} = \sum_{path} P_{path} * P_{good(symb)} * P_{good(x)}$

[0075] $P_{bad} = \sum_{path} P_{path} * P_{bad(symb)} * P_{bad(x)}$

[0076] Если вероятность пути, ведущего в «плохой» кластер, в N раз (1000-10000) выше, чем вероятность пути, ведущего в «хороший» кластер, то эта связь отбрасывается, в противном случае - сохраняется:

[0077] $P_{bad} <> P_{good} * N$

[0078] Программы, используемые для осуществления способов, соответствующих данному изобретению, могут быть частью операционной системы или они могут представлять собой специализированное приложение, компоненту, программу, динамически подключаемую библиотеку, модуль, сценарий или их комбинацию.

[0079] Настоящее описание раскрывает основной изобретательский замысел, который не может быть ограничен упоминавшимися выше аппаратными средствами. Следует отметить, что аппаратные средства в первую очередь предназначены для решения узкой проблемы. С течением времени по мере развития технологии этот тип задач становится более сложным или развивается. Появляются новые инструменты, способные удовлетворять новым требованиям. В этом смысле уместно рассмотреть это оборудование с точки зрения класса технических задач, которые оно способно решить, а не просто как техническую реализацию с использованием некоторого набора компонентов оборудования.

[0080] На Фиг. 5 показан пример вычислительного средства 500, которое может использоваться в настоящем изобретении, которое описано выше. Вычислительное средство 500 включает по меньшей мере один процессор 502, соединенный с памятью 504. Процессор 502 может содержать одно или несколько вычислительных ядер или может представлять собой микросхему или другое устройство, способное выполнять вычисления. Память 504 может представлять собой оперативное запоминающее устройство (ОЗУ), она также может содержать любые другие типы или виды памяти, в частности постоянные запоминающие устройства (например, флэш-накопители) или устройства постоянного хранения данных, такие как жесткие диски, и т.д. Кроме того, можно считать, что память 504 включает аппаратные средства хранения информации, физически расположенные где-либо еще в составе вычислительного средства 500, например, кэш-память в процессоре 502, память, используемую в качестве виртуальной памяти, и сохраняемую во внешнем или внутреннем постоянном запоминающем устройстве 510.

[0081] Вычислительное средство 500 обычно также имеет некоторое количество входов и выходов для передачи информации вовне и получения информации извне. Для взаимодействия с пользователем вычислительное средство 500 может содержать одно или несколько устройств ввода 506 (например, клавиатура, мышь, сканер и т.д.) и устройства вывода 508 (например, дисплеи или специальные индикаторы). Компьютерная система 500 может также иметь одно или несколько постоянных запоминающих устройств 510, такие как оптический привод дисков (формата CD, DVD и др.), Кроме того, вычислительное средство 500 может иметь интерфейс с одной или несколькими сетями 512, которые обеспечивают соединение с другими сетями и вычислительными устройствами. В частности, это может быть локальная сеть (LAN) или беспроводная сеть Wi-Fi, причем она может быть подключена к сети Интернет или не подключена.

[0082] Подразумевается, что вычислительное средство 500 может включать аналоговые и/или цифровые интерфейсы между процессором 502 и каждым из компонентов 504, 506, 508, 510 и 512. Вычислительное средство 500 управляется операционной системой 514, выполняя различные приложения, компоненты, программы, объекты, модули и другое, обозначенное обобщенной цифрой 516.

[0083] Программы, используемые для выполнения способов, соответствующих данному изобретению, могут представлять собой часть операционной системы, либо они могут представлять собой обособленное приложение, компоненту, программу, динамически подключаемую библиотеку, модуль, сценарий или их комбинацию.

[0084] Настоящее описание излагает основной изобретательский замысел авторов, который не может быть ограничен теми аппаратными устройствами, которые упоминались ранее. Следует отметить, что аппаратные устройства, прежде всего, предназначены для решения узкой задачи. С течением времени и с развитием

технического прогресса такая задача усложняется или эволюционирует. Появляются новые средства, которые способны выполнить новые требования. В этом смысле следует рассматривать данные аппаратные устройства с точки зрения класса решаемых ими технических задач, а не чисто технической реализации на некой элеменарной базе.

[0085] Как будет понятно специалистам в данной области, объекты настоящего изобретения могут быть воплощены в виде системы, способа или программного продукта для компьютера. Соответственно, объекты настоящего изобретения могут принимать форму полностью аппаратного варианта осуществления, полностью программного варианта осуществления (в том числе в виде микропрограммного обеспечения, резидентного программного обеспечения, микрокода и т.д.) или варианта осуществления, сочетающего программные и аппаратные объекты, которые могут обобщенно называться в настоящем документе «схема», «модуль» или «система». Кроме того, объекты настоящего изобретения могут иметь вид программного продукта для компьютера, записанного на одном машиночитаемом носителе или нескольких машиночитаемых носителях, содержащих машиночитаемый код записанной на них программы.

[0086] Любая комбинация одного машиночитаемого носителя или нескольких машиночитаемых носителей может быть использована. Машиночитаемый носитель может представлять собой машиночитаемый носитель сигнала или машиночитаемый носитель данных. Например, машиночитаемый носитель данных может представлять собой следующее, но не ограничиваясь этим: электронную, магнитную, оптическую, электромагнитную, инфракрасную, или полупроводниковую систему, аппарат, или устройство, или любую подходящую комбинацию перечисленного выше. Более конкретные примеры машиночитаемых носителей данных (неполный список) включают следующее: электрическое соединение, имеющее один или несколько проводов, портативный компьютерный диск, жесткий диск, запоминающее устройство с произвольным доступом (RAM), постоянное запоминающее устройство (ROM), стираемое программируемое постоянное запоминающее устройство (EPROM или флэш-память), оптоволоконный кабель, портативное постоянное запоминающее устройство на компакт-диске (CD-ROM), оптическое запоминающее устройство, магнитное запоминающее устройство или любую подходящую комбинацию перечисленного выше. В контексте настоящего документа машиночитаемый носитель данных может представлять собой любой материальный носитель, который может содержать или хранить программу для использования системой выполнения команд, аппаратом или устройства или при его соединении с такой системой выполнения команд, аппаратом или устройством.

[0087] Программный код, записанный в машиночитаемом носителе, может передаваться с использованием любой подходящей среды, включая следующие, но не ограничиваясь ими: беспроводная связь, проводная связь, оптический кабель, радиочастотная линия и т.д., или любую подходящую комбинацию перечисленного выше. Программный код для компьютера, выполняющего операции согласно объектам настоящего изобретения, может быть написан на одном или любой комбинации нескольких языков программирования, включая объектно-ориентированные языки программирования, таких как Java, Smalltalk, C++ и т.п., и обычные процедурные языки программирования, такие как язык программирования C или похожие языки программирования. Этот программный код может использоваться полностью на компьютере пользователя, частично на компьютере пользователя в качестве автономного пакета программного обеспечения, частично на компьютере пользователя

и частично на удаленном компьютере или полностью на удаленном компьютере или сервере. В последнем сценарии удаленный компьютер может быть соединен с компьютером пользователя с помощью сети любого типа, в том числе с помощью локальной сети (LAN) или глобальной сети (WAN), или такое соединение может быть выполнено с внешним компьютером (например, с помощью сети Интернет через поставщика услуг Интернет).

[0088] Объекты настоящего изобретения были описаны выше со ссылкой на блок-схемы и/или блок-схемы способов, устройств (систем) и программных продуктов для компьютеров в соответствии с вариантами осуществления настоящего изобретения. Следует понимать, что каждый блок на блок-схеме и комбинации блоков на блок-схемах могут быть реализованы командами программы для компьютера. Эти команды программы могут быть введены в процессор универсального компьютера, специализированного компьютера или другого программируемого устройства обработки данных с тем, чтобы команды, которые выполняются с помощью процессора компьютера или другого программируемого устройства обработки данных, давала возможность осуществления функций или действий, указанных в блок-схеме и/или отдельныом блоке/блоках блок-схемы.

[0089] Эти команды компьютерной программы могут также храниться в машиночитаемом носителе, который может управлять компьютером, другим программируемым устройством обработки данных, или другими устройствами, чтобы они работали определенным образом, так что команды, хранящиеся на машиночитаемом носителе, производили продукт, включая команды, реализующие функцию или действие, предусмотренное в блок-схеме и/или в отдельном блоке/блоках блок-схемы. Эти команды компьютерной программы также могут быть загружены в компьютер, другое программируемое устройство обработки данных или другие устройства для выполнения последовательности рабочих шагов, которые должны осуществляться на компьютере, другом программируемом устройстве или на других устройствах, для обеспечения процесса, реализованного таким образом, чтобы команды, которые выполняются в компьютере или в другом программируемом устройстве, обеспечивали процессы осуществления и выполнения функций или действий, изображенных на блок-схеме и/или в отдельном блоке/блоках блок-схемы.

[0090] Блок-схемы на приведенных выше чертежах иллюстрируют архитектуру, функциональность и работу возможных вариантах осуществления систем, способов и программных продуктов для компьютера в соответствии с различными вариантами осуществления настоящего изобретения. В связи с этим каждый блок на блок-схеме может представлять собой модуль, сегмент или часть кода, которая содержит одну или несколько исполняемых команд для осуществления указанной логической функции (указанных логических функций). Также следует отметить, что в некоторых альтернативных вариантах осуществления функций, показанных в блоке, могут выполняться не в том порядке, который показан на чертежах. Например, два блока, показанные в определенном порядке, фактически в общем случае могут выполняться параллельно, а могут иногда выполняться в обратном порядке, в зависимости от требуемой функциональности. Кроме того, следует отметить, что каждый блок на блок-схеме и/или иллюстрации блок-схемы, и комбинации блоков на блок-схеме и/или иллюстрации блок-схемы, может быть реализован с помощью специализированных систем оборудования, которые выполняют заданные функции или действия, или с помощью комбинации специализированного оборудования и машинных команд.

## Формула изобретения

1. Осуществляемый на компьютере способ для анализа текста, включающий:

идентификацию предложения;

построение графа обобщенных составляющих для предложения на основе грубого синтаксического анализа лексико-морфологической структуры предложения, отличающегося тем, что этот граф обобщенных составляющих содержит дуги и узлы;

фильтрацию дуг графа обобщенных составляющих с использованием комбинированного классификатора с целью сокращения перебора вариантов разбора без потери смысла предложения;

идентификацию наиболее вероятной синтаксической структуры предложения путем осуществления точного синтаксического анализа предложения на основе отфильтрованного графа обобщенных составляющих предложения.

2. Осуществляемый на компьютере способ по п. 1, отличающийся тем, что этот комбинированный классификатор содержит древесный классификатор и по меньшей мере один линейный классификатор.

3. Осуществляемый на компьютере способ по п. 2, отличающийся тем, что древесный классификатор делит дуги на кластеры на основе заранее определенного набора признаков.

4. Осуществляемый на компьютере способ по п. 3, отличающийся тем, что заданный набор признаков основан на параллельном анализе двуязычных текстовых корпусов.

5. Осуществляемый на компьютере способ по п. 3, отличающийся тем, что порядок признаков из предварительно заданного набора признаков определяется на основе оценки энтропии признаков.

6. Осуществляемый на компьютере способ по п. 2, отличающийся тем, что древесный классификатор основан на итерационном дихотомическом алгоритме ID3.

7. Осуществляемый на компьютере способ по п. 2, отличающийся тем, что веса для линейного классификатора основаны на параллельном анализе двуязычных текстовых корпусов.

8. Носители данных для компьютера, содержащие одну или несколько программ для компьютера, отличающиеся тем, что одна или несколько программ для компьютера содержат команды, которые при выполнении устройством обработки данных, приводят к тому, что это устройство обработки данных выполняет операции для анализа текста, включающие:

выявление предложения;

построение графа обобщенных составляющих для предложения на основании грубого синтаксического анализа лексико-морфологической структуры предложения, отличающегося тем, что этот граф обобщенных составляющих содержит дуги и узлы;

фильтрацию дуг графа обобщенных составляющих с использованием комбинированного классификатора с целью сокращения перебора вариантов разбора без потери смысла предложения;

идентификацию наиболее вероятной синтаксической структуры предложения путем осуществления точного синтаксического анализа предложения на основе отфильтрованного графа обобщенных составляющих предложения.

9. Носители данных для компьютера по п. 8, отличающиеся тем, что комбинированный классификатор содержит древесный классификатор и по меньшей мере один линейный классификатор.

10. Носители данных для компьютера по п. 9, отличающиеся тем, что древесный

классификатор разделят дуги по кластерам на основе предварительно определенного набора признаков.

11. Носители данных для компьютера по п. 10, отличающиеся тем, что предварительно заданный набор признаков основан на параллельном анализе двуязычных текстовых корпусов.

12. Носители данных для компьютера по п. 10, отличающиеся тем, что порядок признаков в предварительно заданном наборе признаков определяется на основе оценки энтропии признаков.

13. Носители данных для компьютера по п. 9, отличающиеся тем, что древесный классификатор основан на итерационном дихотомическом алгоритме ID3.

14. Носители данных для компьютера по п. 9, отличающиеся тем, что веса для линейного классификатора основаны на параллельном анализе двуязычных текстовых корпусов.

15. Система для анализа текста, содержащая вычислительное устройство, и машиночитаемый носитель, соединенный с вычислительным устройством и имеющий хранящиеся в нем команды, которые при выполнении вычислительным устройством приводят к тому, что это вычислительное устройство выполняет операции, включающие:

выявление предложения;

построение графа обобщенных составляющих для предложения на основании грубого синтаксического анализа лексико-морфологической структуры предложения, отличающегося тем, что этот граф обобщенных составляющих содержит дуги и узлы;

фильтрацию дуг графа обобщенных составляющих с использованием комбинированного классификатора с целью сокращения перебора вариантов разбора без потери смысла предложения;

идентификацию наиболее вероятной синтаксической структуры предложения путем осуществления точного синтаксического анализа предложения на основе отфильтрованного графа обобщенных составляющих предложения.

16. Система по п. 15, отличающаяся тем, что комбинированный классификатор включает древесный классификатора и по меньшей мере один линейный классификатор.

17. Система по п. 16, отличающаяся тем, что древесный классификатор разделяет дуги на кластеры на основе заранее определенного набора признаков.

18. Система по п. 17, отличающаяся тем, что заданный набор признаков основана на параллельном анализе двуязычных текстовых корпусов.

19. Система по п. 17, отличающаяся тем, что порядок признаков в предварительно заданном наборе признаков определяется на основе оценки энтропии признаков.

20. Система по п. 16, отличающаяся тем, что древесный классификатор основан на итерационном дихотомическом алгоритме ID3.

21. Система по п. 16, отличающаяся тем, что веса для линейного классификатора основаны на параллельном анализе двуязычных текстовых корпусов.

22. Осуществляемый на компьютере способ обучения классификатора, предназначенный для использования в осуществляемом на компьютере способе анализа текста, включающий:

выполнение параллельного анализа неразмеченных двуязычных текстовых корпусов, включающего:

выявление двух соответствующих друг другу предложений в неразмеченных двуязычных текстовых корпусах;

построение двух соответствующих графов обобщенных составляющих для каждого из двух соответствующих предложений;

построение по меньшей мере одного синтаксического дерева на основе графов обобщенных составляющих для двух соответствующих предложений, отличающееся тем, что дуга графа обобщенных составляющих включается в синтаксическое дерево на основании наличия соответствующей дуги у второго графа обобщенных составляющих;

обучение комбинированного классификатора на основе построенного синтаксического дерева.

23. Носители данных для компьютера, предназначенные для использования в осуществляемом на компьютере способе анализа текста, содержащие одну или несколько программ для компьютера, отличающиеся тем, что одна или несколько программ для компьютера содержат команды, при выполнении которых устройство обработки данных выполняет операции, включающие:

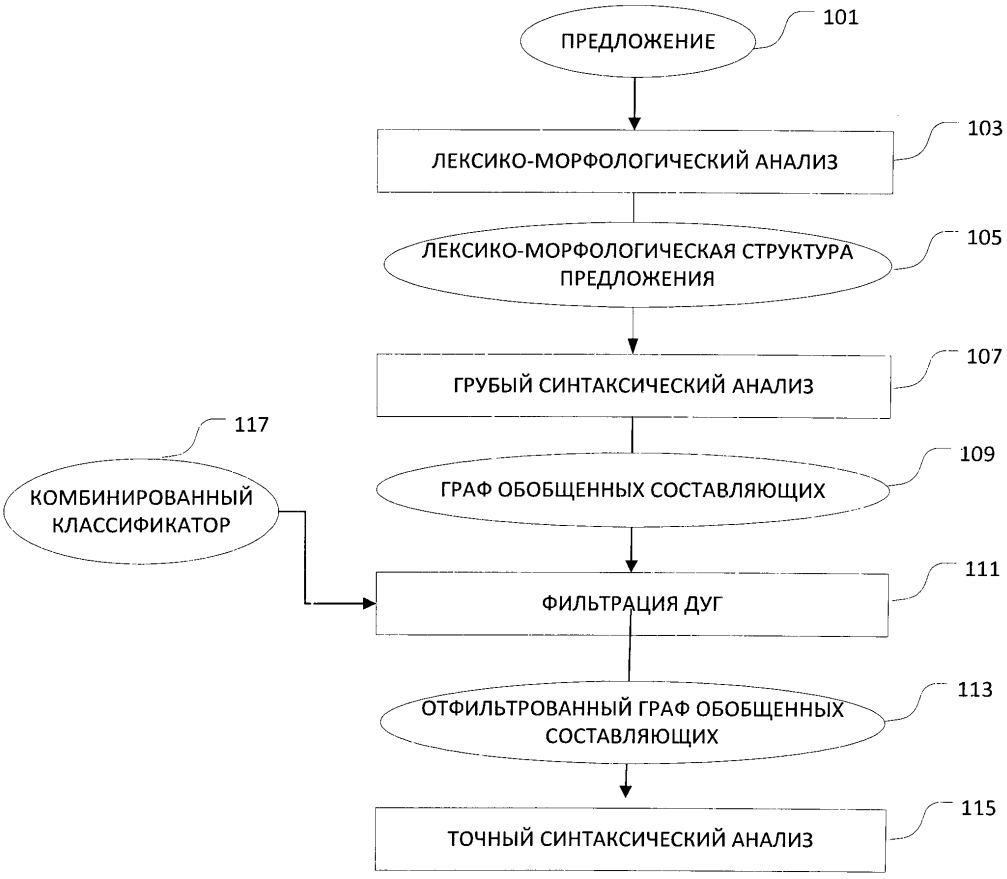
выполнение параллельного анализа неразмеченных двуязычных текстовых корпусов, включающего:

выявление двух соответствующих друг другу предложений в неразмеченных двуязычных текстовых корпусах;

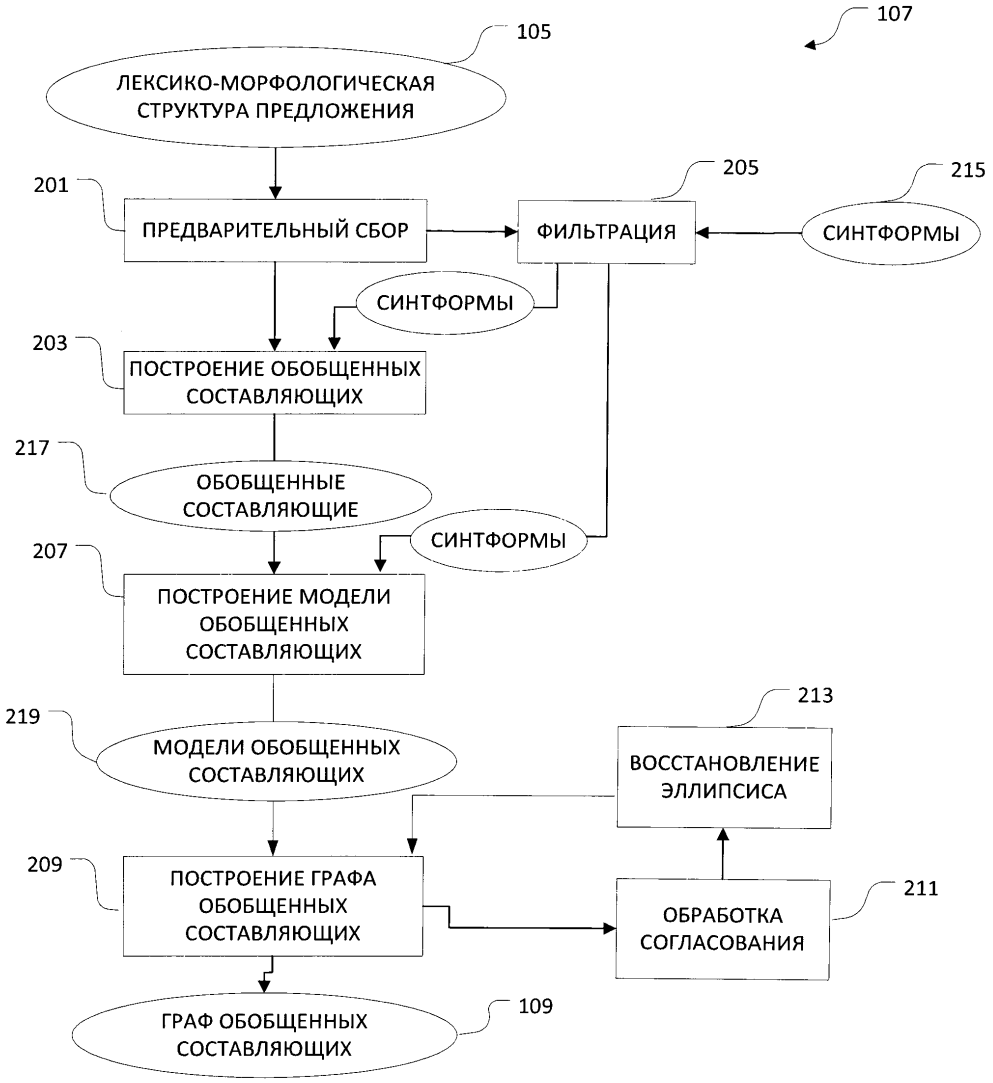
построение двух соответствующих графов обобщенных составляющих для каждого из двух соответствующих предложений;

построение по меньшей мере одного синтаксического дерева на основе графов обобщенных составляющих для двух соответствующих предложений, отличающееся тем, что дуга графа обобщенных составляющих включается в синтаксическое дерево на основании наличия соответствующей дуги у второго графа обобщенных составляющих;

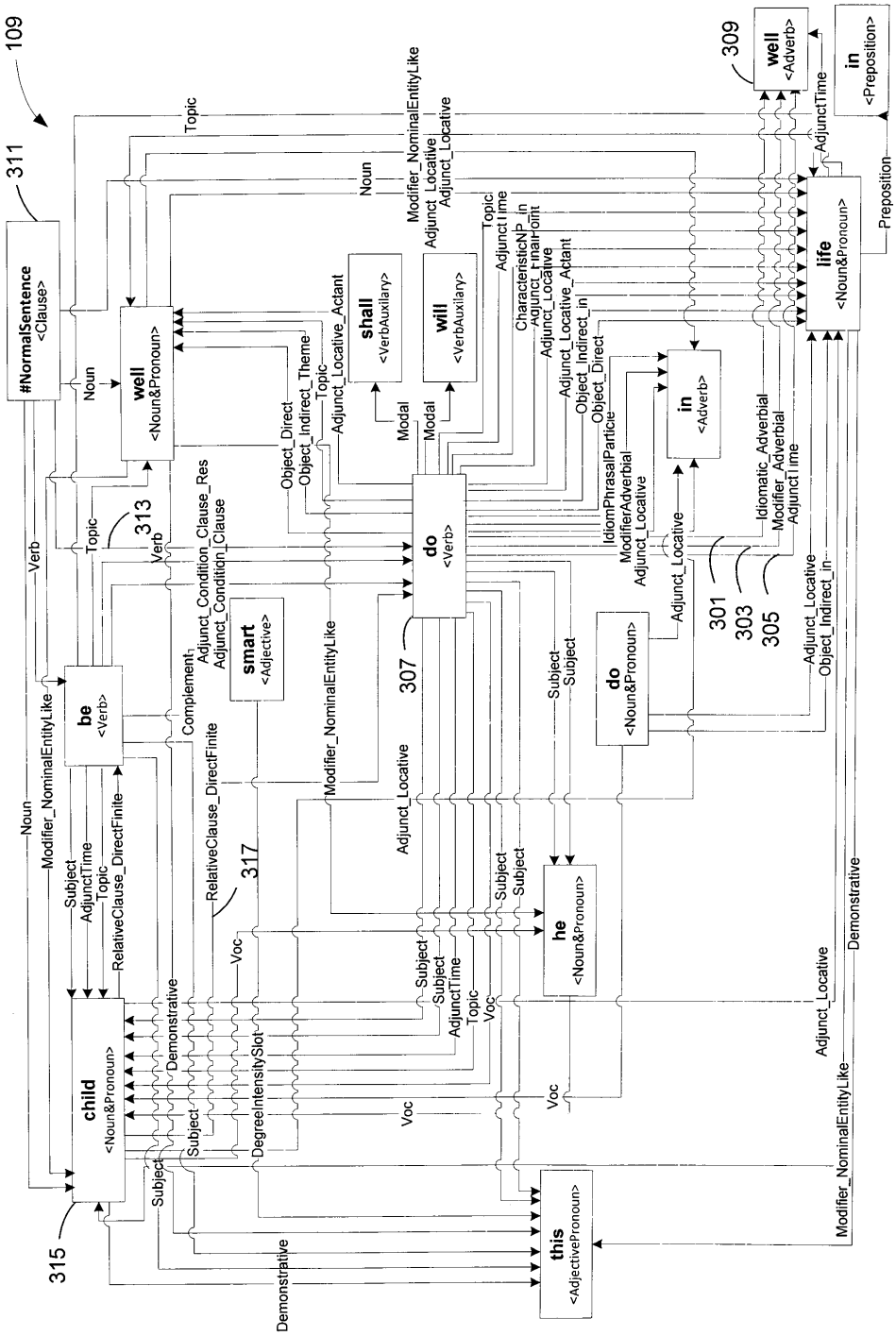
обучение комбинированного классификатора на основе построенного синтаксического дерева.


ПРЕДЛОЖЕНИЕ
101
ЛЕКСИКО-МОРФОЛОГИЧЕСКИЙ АНАЛИЗ
103
ЛЕКСИКО-МОРФОЛОГИЧЕСКАЯ СТРУКТУРА ПРЕДЛОЖЕНИЯ
105
ГРУБЫЙ СИНТАКСИЧЕСКИЙ АНАЛИЗ
107
КОМБИНИРОВАННЫЙ КЛАССИФИКАТОР
117
ГРАФ ОБОБЩЕННЫХ СОСТАВЛЯЮЩИХ
109
ФИЛЬТРАЦИЯ ДУГ
111
ОТФИЛЬТРОВАННЫЙ ГРАФ ОБОБЩЕННЫХ СОСТАВЛЯЮЩИХ
113
ТОЧНЫЙ СИНТАКСИЧЕСКИЙ АНАЛИЗ
115

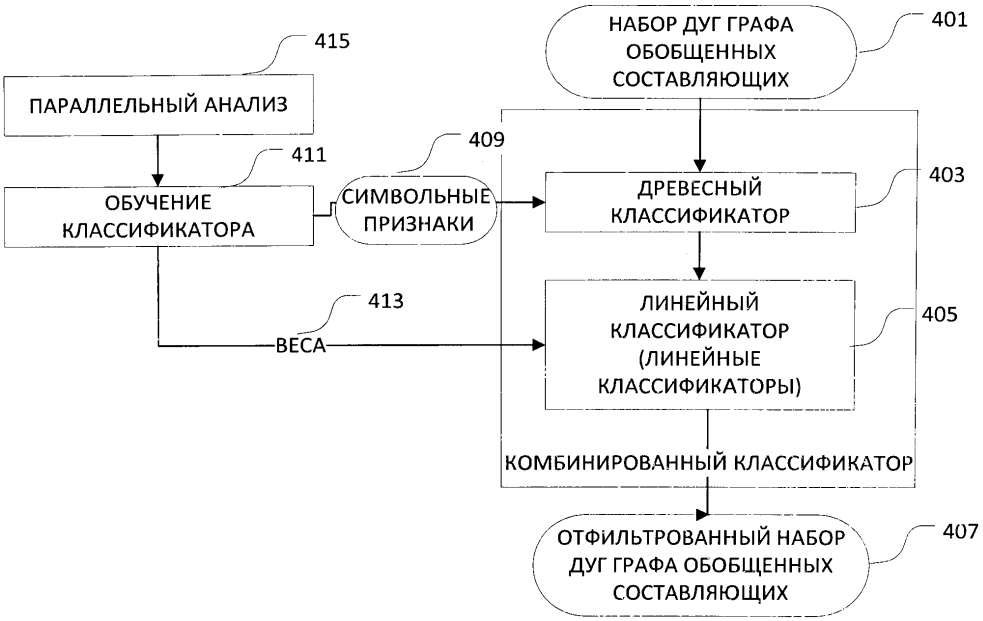
**Фиг. 1**


107
105
ЛЕКСИКО-МОРФОЛОГИЧЕСКАЯ СТРУКТУРА ПРЕДЛОЖЕНИЯ
201
ПРЕДВАРИТЕЛЬНЫЙ СБОР
205
ФИЛЬТРАЦИЯ
215
СИНТФОРМЫ
СИНТФОРМЫ
203
ПОСТРОЕНИЕ ОБОБЩЕННЫХ СОСТАВЛЯЮЩИХ
217
ОБОБЩЕННЫЕ СОСТАВЛЯЮЩИЕ
СИНТФОРМЫ
207
ПОСТРОЕНИЕ МОДЕЛИ ОБОБЩЕННЫХ СОСТАВЛЯЮЩИХ
213
219
МОДЕЛИ ОБОБЩЕННЫХ СОСТАВЛЯЮЩИХ
ВОССТАНОВЛЕНИЕ ЭЛЛИПСИСА
209
ПОСТРОЕНИЕ ГРАФА ОБОБЩЕННЫХ СОСТАВЛЯЮЩИХ
211
ОБРАБОТКА СОГЛАСОВАНИЯ
ГРАФ ОБОБЩЕННЫХ СОСТАВЛЯЮЩИХ
109

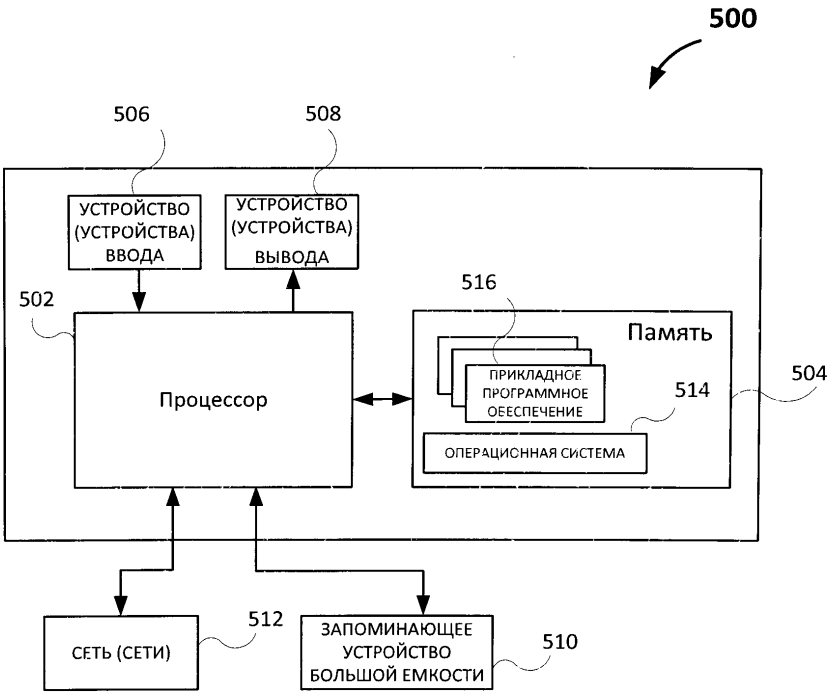
**Фиг. 2**


109
311
#NormalSentence
<Clause>
well
<Noun&Pronoun>
be
<Verb>
child
<Noun&Pronoun>
smart
<Adjective>
shall
<VerbAuxiliary>
will
<VerbAuxiliary>
do
<Verb>
do
<Noun&Pronoun>
he
<Noun&Pronoun>
this
<AdjectivePronoun>
in
<Adverb>
309
well
<Adverb>
in
<Preposition>
life
<Noun&Pronoun>
313
307
315
317
301
303
305
Topic
Noun
Verb
Modifier_NominalEntityLike
Adjunct_Locative
Subject
Voc
Demonstrative
Modal
Object_Direct
Object_Indirect_Theme
Adjunct_Locative_Actant
Adjunct_Condition_Clause_Res
Adjunct_Condition_Clause
Complement
RelativeClause_DirectFinite
DegreeIntensitySlot
AdjunctTime
Preposition
IdiomPhrasalParticle
ModifierAdverbial
Idiomatic_Adverbial
Modifier_Adverbial
Object_Indirect_in

Фиг. 3


НАБОР ДУГ ГРАФА ОБОБЩЕННЫХ СОСТАВЛЯЮЩИХ
401
ПАРАЛЛЕЛЬНЫЙ АНАЛИЗ
415
ОБУЧЕНИЕ КЛАССИФИКАТОРА
411
СИМВОЛЬНЫЕ ПРИЗНАКИ
409
ДРЕВЕСНЫЙ КЛАССИФИКАТОР
403
ВЕСА
413
ЛИНЕЙНЫЙ КЛАССИФИКАТОР (ЛИНЕЙНЫЕ КЛАССИФИКАТОРЫ)
405
КОМБИНИРОВАННЫЙ КЛАССИФИКАТОР
ОТФИЛЬТРОВАННЫЙ НАБОР ДУГ ГРАФА ОБОБЩЕННЫХ СОСТАВЛЯЮЩИХ
407

**Фиг. 4**


500
506
508
УСТРОЙСТВО (УСТРОЙСТВА) ВВОДА
УСТРОЙСТВО (УСТРОЙСТВА) ВЫВОДА
502
516
Память
ПРИКЛАДНОЕ ПРОГРАММНОЕ ОБЕСПЕЧЕНИЕ
514
504
Процессор
ОПЕРАЦИОННАЯ СИСТЕМА
512
СЕТЬ (СЕТИ)
ЗАПОМИНАЮЩЕЕ УСТРОЙСТВО БОЛЬШОЙ ЕМКОСТИ
510

**Фиг. 5**