(19) Евразийское патентное ведомство

(11) 043239 (13) B1

(12) ОПИСАНИЕ ИЗОБРЕТЕНИЯ К ЕВРАЗИЙСКОМУ ПАТЕНТУ

(45) Дата публикации и выдачи патента
2023.04.28

(21) Номер заявки
202192297

(22) Дата подачи заявки
2021.09.16

(51) Int. Cl. ***G06F 40/10*** (2006.01)
G06F 40/20 (2006.01)
G06F 16/35 (2006.01)

(54) СПОСОБ И СИСТЕМА ИЗВЛЕЧЕНИЯ ИМЕНОВАННЫХ СУЩНОСТЕЙ

(31) **2021123151**
(32) **2021.08.03**
(33) **RU**
(43) **2023.02.28**

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

(72) Изобретатель:
Водолазский Даниил Иванович, Гладких Прохор Владимирович, Сорокин Семен Александрович, Черкасов Роман Владиславович, Газизов Куат (RU)

(74) Представитель:
Герасин Б.В. (RU)

(56) US-B2-10007658
US-B2-10489463
US-A1-20180060306
RU-C1-2679988

(57) Изобретение в общем относится к области вычислительной техники, а в частности к способу и системе извлечения именованных сущностей из текстовой информации на основе двухуровневой классификации именованных сущностей. Техническим результатом от реализации заявленного способа является повышение точности распознавания именованных сущностей, за счет двухуровневой классификации. Указанный технический результат достигается благодаря осуществлению компьютерно-реализуемого способа извлечения именованных сущностей из текстовой информации с помощью модели машинного обучения, выполняемого с помощью по меньшей мере одного процессора и содержащего этапы, на которых получают входные текстовые данные и выполняют их обработку, причем обработка включает по меньшей мере сегментацию и токенизацию сегментированных данных; выполняют векторизацию полученных токенов; осуществляют обработку полученных данных с помощью модели машинного обучения на базе нейронной сети, извлекают по меньшей мере одну именованную сущность с определенным по меньшей мере одним подтипом класса.

043239 B1

043239 B1

Область техники

Изобретение в общем относится к области вычислительной техники, а в частности к способу и системе извлечения именованных сущностей из текстовой информации на основе двухуровневой классификации именованных сущностей.

Уровень техники

Задача распознавания именованных сущностей (Named-entity recognition, NER) является одной из самых популярных задач обработки текстов на естественном языке (Natural Language Processing, NLP), а также применяется в большом количестве приложений автоматической обработки текстов и информационного поиска. Так, NER может быть использована для извлечения из новостных источников имен людей, названий организаций, географических мест и др. Также задача NER активно применяется при построении диалоговых систем (чат-ботов, умных помощников). Для распознавания именованных сущностей используются подходы, основанные на выделении в тексте последовательностей слов, являющихся именованными сущностями, и классификации выделенных именованных сущностей. Примерами классов именованных сущностей являются имена людей, названий организаций, географических названий, прочие типы имен собственных, а также выражения специального вида, такие как обозначения моментов времени, дат, денежные суммы и процентные выражения.

Однако несмотря на востребованность методов решения задачи распознавания именованных сущностей, в настоящее время существуют трудности с распознаванием именованных сущностей, которые могут пересекаться между собой и при этом обладать атрибутами (подклассами/подтипами), которые, в свою очередь, допускают единственный или множественный выбор. Такие проблемы, как правило, сводятся к задаче многоклассовой классификации (если подтип всегда принимает ровно одно значение, multiclass classification) либо многометочной классификации (если подтип может принимать одно, несколько или ни одного значения, multilabel classification). Например, если именованная сущность - название фильма, то многоклассовым подтипом может быть возрастной рейтинг, а многометочным - жанры. Кроме того, важной областью, где требуется извлечение сложных сущностей, например, имеющих несколько подтипов сущности, пересекающихся сущностей, вложенных сущностей и т.д., является обработка юридических документов, например, судебных решений. Поэтому извлечение сложных именованных сущностей является существенной задачей.

Из уровня техники известен способ классификации документов, раскрытый в патенте США № US 10552501 B2 (OATH INC), опубл. 04.02.2020. Указанный способ обеспечивает возможность классификации нового документа одновременно несколькими метками (Multilabel classification) за счет представления размеченных документов в векторном пространстве таким образом, чтобы документы и связанные с ними метки располагались в непосредственной близости друг от друга. Для классификации нового документа указанный документ представляется в векторном пространстве и выполняется поиск близости (например, методом K-ближайших соседей) в векторном пространстве для идентификации набора ближайших меток к новому документу.

К недостаткам указанного способа можно отнести невозможность применения указанного способа в задачах NER, а также невозможность выполнения двухуровневой классификации данных, основанной на определении подтипов класса именованной сущности.

Из уровня техники известен подход к распознаванию именованных сущностей от DeepPavlov, URL: http://docs.deeppavlov.ai/en/master/features/models/ner.html. Указанная модель классифицирует слова текста по типу именованных сущностей, к которым они принадлежат.

Недостатками указанного решения являются отсутствие поддержки пересекающихся сущностей и классификации подтипов классов именованных сущностей, что влияет на точность классификации моделью машинного обучения именованных сущностей и, как следствие, не обеспечивает высокого качества при применении модели машинного обучения для извлечения именованных сущностей.

Общими недостатками существующих решений является отсутствие эффективного способа извлечения именованных сущностей из текстовой информации, обеспечивающего возможность классификации именованных сущностей на подтипы класса, к которому принадлежит указанная сущность, и извлечения пересекающихся сущностей. Кроме того, такого рода решение должно обеспечивать возможность распознавания именованных сущностей с высоким качеством за счет повышения точности классификации именованных сущностей.

Раскрытие изобретения

В заявленном техническом решении предлагается новый подход к распознаванию именованных сущностей в текстовой информации. В данном решении используется алгоритм машинного обучения, который позволяет осуществлять двухуровневую классификацию именованных сущностей, что повышает точность их распознавания и качество применения модели машинного обучения для извлечения именованных сущностей.

Таким образом, решается техническая проблема обеспечения возможности извлечения пересекающихся именованных сущностей и классификации подтипов класса именованных сущностей.

Техническим результатом, достигающимся при решении данной проблемы, является повышение точности распознавания именованных сущностей за счет двухуровневой классификации.

Дополнительным техническим результатом, проявляющимся при решении вышеуказанной проблемы, является обеспечение возможности извлечения пересекающихся и вложенных сущностей.

Указанные технические результаты достигаются благодаря осуществлению компьютерно-реализуемого способа извлечения именованных сущностей из текстовой информации с помощью модели машинного обучения, выполняемый с помощью по меньшей мере одного процессора и содержащий этапы, на которых:

a) получают входные текстовые данные и выполняют их обработку, причем обработка включает по меньшей мере сегментацию и токенизацию сегментированных данных;

b) выполняют векторизацию полученных токенов;

c) осуществляют обработку полученных данных с помощью модели машинного обучения на базе нейронной сети, причем в ходе указанной обработки осуществляется:

преобразование вектора каждого токена, полученного на этапе b), в вектор вероятностей принадлежности соответствующего токена заданному классу;

преобразование вектора токена, полученного на этапе b), в вектор логарифмов вероятностей принадлежности соответствующего токена по меньшей мере одному подтипу класса из заданных подтипов класса;

объединение подряд идущих токенов с одинаковым классом по меньшей мере в одну именованную сущность;

определение принадлежности по меньшей мере одной именованной сущности по меньшей мере одному подтипу класса, причем определение выполняется с помощью этапа, при котором:

векторы логарифмов вероятностей принадлежности токенов по меньшей мере одному подтипу класса суммируются и сравниваются с заданным пороговым значением (при определении подтипа класса, имеющего множество подтипов); или

векторы логарифмов вероятностей принадлежности токенов по меньшей мере одному подтипу класса суммируются и сравниваются между собой, причем в ходе сравнения выбирается подтип, имеющий максимальную вероятность (при определении подтипа класса, имеющего только один подтип);

d) извлекают по меньшей мере одну именованную сущность с определенным по меньшей мере одним подтипом класса.

В одном из частных вариантов реализации способа сегментация представляет собой разбиение текстовых данных по меньшей мере на одно предложение.

В другом частном варианте реализации способа токенизация сегментированных входных текстовых данных содержит этап токенизации данных на подслова.

В другом частном варианте реализации способа токенизация данных на подслова выполняется по меньшей мере с использованием алгоритма токенизации, выбираемого из группы: алгоритм BPE, алгоритм WordPiece.

В другом частном варианте реализации способа вектор каждого токена представляет собой вектор фиксированного размера.

В другом частном варианте реализации способ дополнительно содержит этап отображения по меньшей мере одной извлеченной сущности.

Кроме того, заявленные технические результаты достигаются за счет системы извлечения именованных сущностей из текстовой информации, содержащей:

по меньшей мере один процессор;

по меньшей мере одну память, соединенную с процессором, которая содержит машиночитаемые инструкции, которые при их выполнении по меньшей мере одним процессором обеспечивают выполнение способа извлечения именованных сущностей из текстовой информации.

Краткое описание фигур

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

Фиг. 1 иллюстрирует блок-схему выполнения заявленного способа.

Фиг. 2 иллюстрирует блок-схему выполнения алгоритма двухуровневой классификации именованных сущностей.

Фиг. 3 иллюстрирует архитектуру модели машинного обучения, предназначенную для двухуровневой классификации именованных сущностей.

Фиг. 4 иллюстрирует пример общего вида вычислительной системы, которая обеспечивает реализацию заявленного решения.

Осуществление изобретения

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

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

F1-мера представляет собой совместную оценку точности и полноты.

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

Многометочная классификация (мультилейбл классификация, multilabel classification) - задача о классификации подтипа, принимающего одно, несколько или ни одного значения. Указанная задача широко известна из уровня техники и более подробно раскрыта, например, в источнике https://en.wikipedia.org/wiki/Multi-label_classification, найдено 11.06.2021.

Многоклассовая классификация (multiclass classification) - задача о классификации подтипа, принимающего всегда ровно одно значение. Указанная задача широко известна из уровня техники и более подробно раскрыта, например, в источнике https://en.wikipedia.org/wiki/Multiclass_classification, найдено 11.06.2021.

Заявленное техническое решение предлагает новый подход, обеспечивающий повышение точности распознавания и извлечения именованных сущностей из текстовой информации за счет двухуровневой классификации именованных сущностей. Основной особенностью заявленного технического решения является возможность распознавания пересекающихся именованных сущностей, а также подтипов класса именованных сущностей, что, в свою очередь, повышает точность распознавания именованной сущности и делает возможным применение указанного решения, в частности, для обработки юридических документов.

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

В данном решении под системой подразумевается компьютерная система, ЭВМ (электронно-вычислительная машина), ЧПУ (числовое программное управление), ПЛК (программируемый логический контроллер), компьютеризированные системы управления и любые другие устройства, способные выполнять заданную, чётко определённую последовательность вычислительных операций (действий, инструкций). Под устройством обработки команд подразумевается электронный блок либо интегральная схема (микропроцессор), исполняющая машинные инструкции (программы)/ Устройство обработки команд считывает и выполняет машинные инструкции (программы) с одного или более устройства хранения данных, например таких устройств, как оперативно запоминающие устройства (ОЗУ) и/или постоянные запоминающие устройства (ПЗУ). В качестве ПЗУ могут выступать, но, не ограничиваясь, жесткие диски (HDD), флэш-память, твердотельные накопители (SSD), оптические носители данных (CD, DVD, BD, MD и т.п.) и др.

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

Термин "инструкции", используемый в этой заявке, может относиться, в общем, к программным инструкциям или программным командам, которые написаны на заданном языке программирования для осуществления конкретной функции, такой как, например, получение артефактов программно-аппаратного решения, формирование цифрового стандарта программно-аппаратного решения, формирование результатов проверки программно-аппаратного решения, анализ данных и т. п. Инструкции могут быть осуществлены множеством способов, включающих в себя, например, объектно-ориентированные методы. Например, инструкции могут быть реализованы, посредством языка программирования C++, Java, Python, различных библиотек (например, MFC; Microsoft Foundation Classes) и т. д. Инструкции, осуществляющие процессы, описанные в этом решении, могут передаваться как по проводным, так и по беспроводным каналам передачи данных, например Wi-Fi, Bluetooth, USB, WLAN, LAN и т. п. Обучение модели МО производилось на заранее размеченных данных. На момент создания модели был использован датасет (набор данных) из 1700 размеченных судебных актов. Обучение модели двухуровневой NER заключалось в распознавании основных заданных классов именованных сущностей в тексте (дата, судебное решение, участник, сумма и т.д.) и дальнейшем определении подтипов класса для выявленных именованных сущностей.

Основная метрика качества, взвешенная по классам f-1 мера, составляет около 0.8. Ниже приведены метрики на резолютивной части решения суда. Резолютивная часть решения суда - это часть, которая содержит выводы суда об удовлетворении иска либо об отказе в удовлетворении иска полностью или в части, указание на распределение судебных расходов, срок и порядок обжалования решения суда. Резолютивная часть решения суда должна содержать выводы суда об удовлетворении иска либо об отказе в удовлетворении иска полностью или в части, указание на распределение судебных расходов, срок и порядок обжалования решения суда.

Метрики f-1 меры для различных классифицированных именованных сущностей, извлеченных из резолютивной части судебных решений, полученные в результате обучения модели машинного обучения двухуровневой NER:

Класс сущности: Дата.

Подтипы класса сущности:
Дата объявления полного текста 0.92.
Дата объявления текста резолютивной части 0.84.
Дата предыдущего определения 0.88.
Дата предыдущего постановления 0.97.
Класс сущности: Судебное решение.
Подтипы класса сущности:
Решение о прекращении производства 0.95.
Решение об отмене требований о включении 0.85.
Решение о включении требований в реестр 0.85.
Решение об оставлении требовании без судебного процесса 0.86.
Класс сущности: Сумма.
Подтипы класса сущности:
Сумма основного долга 0.89.
Проценты за пользование средствами 0.92.
Сумма штрафа 0.88.
Сумма пени 0.96.
Класс сущности: Очередь.
Подтипы класса сущности:
1-я очередь 1.0.
2-я очередь 0.85.
3-я очередь 0.83.
Класс сущности: Участник.
Подтипы класса сущности:
Должник 0.89.
Кредитор 0.90.
Новый кредитор 0.80.

На фиг. 1 представлена блок схема способа 100 извлечения именованных сущностей из текстовой информации, который раскрыт поэтапно более подробно ниже. Указанный способ 100 заключается в выполнении этапов, направленных на обработку различных цифровых данных. Обработка, как правило, выполняется с помощью системы, которая может представлять, например, сервер, компьютер, мобильное устройство, вычислительное устройство и т.д. Более подробно элементы системы раскрыты на фиг. 4. На этапе 110 получают входные текстовые данные и выполняют их обработку, причем обработка включает по меньшей мере сегментацию и токенизацию сегментированных данных.

На указанном этапе 110 система 400 получает входные текстовые данные, из которых требуется извлечь необходимую информацию. Входные текстовые данные могут представлять собой, например, судебные решения, договоры купли продажи, аннотации, и т. д., не ограничиваясь. Входные текстовые данные могут быть получены, например, посредством загрузки через интерфейс ввода-вывода в систему 400 и/или посредством сети связи (как по беспроводному, так и по проводному соединению). В одном частном варианте осуществления входные текстовые данные могут быть получены по меньшей мере из одной базы данных (БД), в которой содержится текстовая информация. В другом частном варианте осуществления входные текстовые данные могут быть получены от устройства пользователя, например портативного или стационарного компьютера, мобильного телефона или смартфона, планшета и т.д., не ограничиваясь.

Далее выполняется обработка полученных текстовых данных. Входной текст сегментируется на равные отрезки, например, на предложения, предложения на слова и тонизируется. Входной текст может сегментироваться, например, на предложения, а каждый сегмент далее может быть разделен на токены. Под токеном в данном решении следует понимать последовательность символов в документе, которая имеет значение для анализа. Так, токенами могут являться, например, отдельные слова, части слов и т.д. Сегментация на предложения может проводиться при помощи, например, лексических анализаторов, таких как razdel, rusenttokenize, NLTK и т.д. Лексический анализ - это процесс аналитического разбора входной последовательности символов на распознанные группы (лексемы) с целью получения на выходе идентифицированных последовательностей, называемых "токенами". Кроме того, сегментация входного текста может быть осуществлена на основе регулярных выражений. Процесс токенизации также может быть выполнен при помощи лексических анализаторов, которые были описаны выше. Для специалиста в данной области техники очевидно, что может быть применен любой лексический анализатор известный из уровня техники и данное решение не ограничивается приведенными выше примерами.

В еще одном частном варианте осуществления токенизация может представлять собой разбиение текста на слова по пробелу между словами и токенизацию каждого слова на подслова с помощью алгоритма WordPiece для получения последовательности токенов. Например, для текста "Должник обязан погасить задолженность" последовательность токенов будет представлять: Дол', '##жник', 'о', '##бя', '##зан', 'по', '##гас', '##ить', 'за', '##дол' '##жен' '##ность'. В другом частном варианте осуществления токе-

низация текста может быть выполнена с помощью алгоритма BPE (Byte Pair encoding). Указанный алгоритм также разбивает текст на подслова, как показано выше. Сформированный набор токенов далее образует словарь фиксированного размера (например, 30000 токенов). Для этого часто встречающиеся группы символов не заменяются на другой символ, а объединяются в токен и добавляются в словарь. За счет этого алгоритм токенизации на основе BPE позволяет распознавать как можно больше слов при ограниченном объеме словаря. На этапе 120 выполняют векторизацию полученных токенов. На указанном этапе 120 выполняется векторизация каждого токена, полученного в процессе токенизации, например, с помощью прямого кодирования (one hot encoding). Так, например, при токенизации на основе алгоритма BPE, каждый токен, полученный в ходе указанного процесса токенизации, представлен в словаре своим индексом, отображающий позицию в указанном словаре. Таким образом, каждый токен представляет бинарный вектор (значения 0 или 1), а единица ставится тому элементу, который соответствует номеру токена в словаре, что позволяет представить каждый токен в виде вектора фиксированной длины, соответствующей размерности словаря. Для специалиста в данной области техники будет очевидно, что для векторизации токенов могут применять и другие алгоритмы векторизации, например, с помощью алгоритмов Word2vec, fastText и т.д., не ограничиваясь.

Далее способ 100 переходит к этапу 130. На этапе 130 осуществляют обработку полученных данных с помощью модели машинного обучения на базе нейронной сети. Как указывалось выше, модель машинного обучения была обучена на заранее размеченных данных. Метрики качества модели были приведены выше.

В ходе обработки данных с помощью модели машинного обучения выполняется следующая последовательность шагов алгоритма двухуровневой классификации именованных сущностей 200, который отображен на фиг. 2. Указанные шаги выполняются с помощью модели машинного обучения, построенной на базе нейронной сети. В качестве нейронной сети может быть использована, например, нейронная сеть трансформер (Transformer), рекуррентная нейронная сеть и т.д., не ограничиваясь. Особенностью указанной модели машинного обучения является возможность эффективного распознавания пересекающихся и вложенных именованных сущностей за счет их двухуровневой классификации. Так, указанное техническое решение позволяет осуществлять многоклассовую классификацию (если подтип всегда принимает ровно одно значение, multiclass classification) либо многометочную классификацию (если подтип может принимать одно, несколько или ни одного значения, multilabel classification). Кроме того, за счет повышения точности классификации моделью машинного обучения именованных сущностей, как следствие, обеспечивается высокое качество извлечения именованных сущностей.

Так, например, за счет реализации настоящей модели машинного обучения для двухуровневой NER (двухуровневая классификация именованных сущностей) обеспечивается возможность распознавания именованных сущностей в судебных решениях. Сложность такой задачи, как было описано выше, заключается в том, что такие решения могут содержать пересекающиеся и вложенные сущности, а распределение по подтипам класса может являться несбалансированным. Например, когда на 1000 сущностей типа "Решение" есть только 7 с подтипом "Отменить старое решение, включить в реестр во вторую очередь", выучить сущность со сложным типом "Решение. Отменить старое решение, включить в реестр во вторую очередь" обычная модель МО (машинное обучение) не сможет. Модель машинного обучения для двухуровневой NER, в свою очередь, осуществляет классификацию сущности на более крупный тип (класс), например, "Решение", точность распознавания которого гораздо выше, чем у упомянутой сущности со сложным типом, и далее выбирает подтип для указанного распознанного класса, например, один из пятнадцати подтипов, что упрощает задачу для модели, повышает точность классификации и, как следствие, повышает качество извлечения именованных сущностей.

Возвращаясь к алгоритму 200, на этапе 210 выполняют преобразование вектора каждого токена, полученного в ходе векторизации, в вектор вероятностей принадлежности соответствующего токена заданному классу. На указанном этапе 210 вектор каждого токена подается в первый классификатор. В качестве классификатора может быть использован, например, многослойный перцептрон с сигмоидной функцией активации. В другом частном варианте осуществления в качестве классификатора может быть использован многослойный перцептрон с функцией активации softmax (софтмакс), например, при многоклассовой, а не мультиметочной классификации. Функция активации softmax широко известна из уровня техники (см., например, https://ru.wikipedia.org/wiki/Softmax). Результатом применения этого классификатора к вектору является вектор вероятностей принадлежности соответствующего токена заданному классу. Как упоминалось выше, количество классов и их нумерация задаются заранее перед обучением модели, в зависимости от требуемой задачи.

На этапе 220 выполняют преобразование вектора токена, полученного в ходе векторизации, в вектор логарифмов вероятностей принадлежности соответствующего токена по меньшей мере одному подтипу класса из заданных подтипов класса. На этапе 220 вектор токена также подается на второй классификатор. Стоит отметить, что этап 210 и этап 220 могут выполняться параллельно и независимо друг от друга. Второй классификатор может представлять собой, например, многослойный перцептрон с сигмоидной функцией активации. Результатом применения второго классификатора к вектору будет являться вектор логарифмов вероятностей принадлежности соответствующего токена по меньшей мере одному

подтипу класса из заданных подтипов класса.

На этапе 230 объединяют подряд идущие токены с одинаковым классом по меньшей мере в одну именованную сущность. На указанном этапе 230 подряд идущие токены объединяются в спаны. Под спаном в данном решении следует понимать непрерывный фрагмент текста (интервал слов). Поскольку именованные сущности - это объекты определенного типа, чаще всего составные, для предсказания класса и подтипов всей сущности токены объединяются в замкнутый интервал слов указанной сущности. В одном частном варианте осуществления подряд идущие токены с одинаковым классом объединяются в ровно одну именованную сущность (спан сущности). В другом частном варианте осуществления, например, в случае использования схемы разметки BIO (раскрыта более подробно ниже), последовательность подряд идущих токенов с одинаковым классом может быть разбита на несколько именованных сущностей, например, на два спана сущностей. Для специалиста в данной области техники очевидно, что объединение подряд идущих токенов с одинаковым классом по меньшей мере в одну именованную сущность буде зависеть от схемы разметки и указанные токены могут быть объединены как в одну именованную сущность, так и в множество именованных сущностей (например, две именованные сущности).

На этапе 240 определяют принадлежность по меньшей мере одной именованной сущности по меньшей мере одному подтипу класса, причем определение выполняется с помощью этапа, при котором: векторы логарифмов вероятностей принадлежности токенов по меньшей мере одному подтипу класса суммируются и сравниваются с заданным пороговым значением (при определении подтипа класса, имеющего множество подтипов); или векторы логарифмов вероятностей принадлежности токенов по меньшей мере одному подтипу класса суммируются и сравниваются между собой, причем в ходе сравнения выбирается подтип имеющий максимальную вероятность (при определении подтипа класса, имеющего только один подтип).

Указанный этап может выполняться вторым и третьим классификатором. Третий классификатор повторяет второй во всём, кроме функции активации (softmax вместо сигмоиды) и числа классов. Третий классификатор отвечает за типы, имеющие ровно одно значение подтипа. Таким образом, на указанном этапе 240 модель машинного обучения определяет либо подтип именованной сущности для класса, имеющего выбор только одного подтипа из нескольких, с помощью третьего классификатора на основе функции активации softmax за счет выбора подтипа класса, имеющего максимальную вероятность среди всех подтипов этого класса, либо определяет подтипы класса именованной сущности для класса, имеющего множественный выбор подтипов, с помощью второго классификатора, описанного выше. Причем, для определения нескольких подтипов класса вектора логарифмов вероятностей принадлежности токенов по меньшей мере одному подтипу класса суммируются и сравниваются с заданным пороговым значением. Пороговое значение может быть задано на стадии обучения модели, а также может быть задано вручную, в зависимости от требуемой точности определения подтипов класса. Верхнеуровневая архитектура указанной модели представлена на фиг. 3.

На фиг. 3 представлена модель машинного обучения 300 для двухуровневой классификации именованных сущностей построенная на базе модели глубокого обучения 310, содержащей классификаторы 320 и 330. В одном частном варианте осуществления в качестве модели 310 может быть применена State-of-the-Art-модель на базе BERT, однако для специалиста в данной области техники очевидно, что может применяться любая модель машинного обучения. Классификаторы 320 и 330 могут являться как программными, так и программно-аппаратными модулями, реализованными на базе системы 400. В качестве наглядного примера работы указанной модели 300 на вход модели была подана следующая текстовая информация, сегментированная на токены 311: "Asia Bibi pochodzi z Pakistanu". Входные токены 311 далее дополнительно токенизируются с помощью алгоритма BPE, как было описано выше, на токены 312. Токены 312 далее подаются в эмбеддер для векторизации. Выходом эмбеддера являются векторизированные токены 313, которые поступают в энкодер. Энкодер для каждого входного токена возвращает вектор вещественных чисел фиксированного размера 314, определяемого архитектурой модели. Далее токены размечаются классификатором 315 в соответствии со схемой BIO. Схема заключается в том, чтобы к метке сущности (например, PER для персон или ORG для организаций) добавить некоторый префикс, который обозначает позицию токена в спане сущности. Более подробно:

B - от слова beginning - первый токен в спане сущности,

I - от слова inside - это то, что находится в середине.

Если токен не относится ни к какой сущности, он помечается специальной меткой, обычно имеющей обозначение OUT или O. Таким образом, размеченные токены 315 далее подаются на вход классификаторам 320 и 330. Для специалиста в данной области техники очевидно, что помимо схемы BIO могут использоваться и другие схемы разметки, например, схема 10 и т.д., не ограничиваясь. Как упоминалось выше, в зависимости от выбираемой схемы разметки подряд идущие токены с одинаковым классом могут быть объединены по меньшей мере в одну именованную сущность (спан сущности). Так, например, в случае разметки BIO, токены могут быть объединены в два спана сущности. Рассмотрим пример: последовательность B - PER, I - PER, B - PER будет преобразована в два спана сущности, где первой именованной сущностью будет являться первые два подряд идущих токена с одинаковым классом, а второй именованной сущностью будет являться третий токен. Классификатор 320 классифицирует подтипы,

когда возможен выбор ровно одного подтипа. Для PER (персоны) это пол (мужской, женский), для LOC (локации) это географическая характеристика (город, страна, континент, море).

Классификатор 330 классифицирует подтипы, когда возможен выбор одного, нескольких подтипов или ни одного. Для ORG (организации) это сферы деятельности.

Как видно на фиг. 3, на каждый токен классификатор выдает предсказания вектора логарифмов вероятностей принадлежности токенов по меньшей мере одному подтипу класса, а затем агрегирует их по всем токенам с одинаковыми классами, объединенными в сущности (спан сущности). На основании вероятностей в агрегированных векторах выбираются итоговые подтипы, как было описано выше (этап 240). В данном примере для токенов "Asia" и "Bibi" максимальную вероятность имеют классы B-PER и I-PER соответственно, где B и I позиция токена в спане сущности. Они объединяются в спан PER из двух подряд идущих токенов. Для каждого из них вычисляются вероятности (логарифмы вероятностей) принадлежности к подтипам класса PER, в данном случае пол (мужской, женский). После агрегации векторов вероятностей по токенам спана выбирается подтип сущности (пол - женский).

Таким образом, на этапе 130 с помощью модели машинного обучения 300 выполняется двухуровневая классификация именованных сущностей.

Далее способ 100 переходит к этапу 140. На этапе 140 извлекают по меньшей мере одну именованную сущность с определенным по меньшей мере одним подтипом класса.

Данные о классифицированных именованных сущностях далее могут быть направлены в хранилище данных для их последующего отображения пользователю по соответствующему запросу, например, с помощью интерфейса ввода-вывода системы 400.

Кроме того, в еще одном частном варианте осуществления классифицированные именованные сущности могут быть представлены в виде таблицы, содержащий сущность и значение сущности.

Таким образом, за счет двухуровневой классификации сущностей, содержащих как множественное значение подтипов класса, так и одно значение подтипов класса, повышается точность распознавания именованных сущностей текстовой информации.

Кроме того, указанное решение также позволяет извлекать именованные сущности из сложных юридических документов за счет возможности распознавания пересекающихся сущностей и классификации подтипов класса сущности.

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

В общем случае система (400) содержит такие компоненты, как: один или более процессоров (401), по меньшей мере одну память (402), средство хранения данных (403), интерфейсы ввода/вывода (404), средство В/В (405), средство сетевого взаимодействия (406), которые объединяются посредством универсальной шины.

Процессор (401) выполняет основные вычислительные операции, необходимые для обработки данных при выполнении способа (100). Процессор (401) исполняет необходимые машиночитаемые команды, содержащиеся в оперативной памяти (202). Память (402), как правило, выполнена в виде ОЗУ и содержит необходимую программную логику, обеспечивающую требуемый функционал.

Средство хранения данных (403) может выполняться в виде HDD, SSD дисков, рейд массива, флэш-памяти, оптических накопителей информации (CD, DVD, MD, Blue-Ray дисков) и т.п. Средства (403) позволяют выполнять долгосрочное хранение различного вида информации, например истории обработки транзакционных запросов (логов), идентификаторов пользователей и т. п.

Для организации работы компонентов системы (400) и организации работы внешних подключаемых устройств применяются различные виды интерфейсов В/В (404). Выбор соответствующих интерфейсов зависит от конкретного исполнения вычислительного устройства, которые могут представлять собой, не ограничиваясь: PCI, AGP, PS/2, IrDa, FireWire, LPT, COM, SATA, IDE, Lightning, USB (2.0, 3.0, 3.1, micro, mini, type C), TRS/Audio jack (2.5, 3.5, 6.35), HDMI, DVI, VGA, Display Port, RJ45, RS232 и т.п. Выбор интерфейсов (404) зависит от конкретного исполнения системы (400). которая может быть реализована на базе широко класса устройств, например, персональный компьютер, мейнфрейм. ноутбук, серверный кластер, тонкий клиент, смартфон, сервер и т. п.

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

граммно-аппаратного решения) по проводному или беспроводному каналу передачи данных, например, WAN, PAN, ЛВС (LAN), Интранет, Интернет, WLAN, WMAN или GSM.

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

Представленные материалы заявки раскрывают предпочтительные примеры реализации технического решения и не должны трактоваться как ограничивающие иные, частные примеры его воплощения, не выходящие за пределы испрашиваемой правовой охраны, которые являются очевидными для специалистов соответствующей области техники. Таким образом, объем настоящего технического решения ограничен только объемом прилагаемой формулы.

ФОРМУЛА ИЗОБРЕТЕНИЯ

1. Компьютерно-реализуемый способ извлечения именованных сущностей из текстовой информации с помощью модели машинного обучения, выполняемый с помощью по меньшей мере одного процессора и содержащий этапы, на которых:

a) получают входные текстовые данные и выполняют их обработку, причем обработка включает по меньшей мере сегментацию и токенизацию сегментированных данных;

b) выполняют векторизацию полученных токенов;

c) осуществляют обработку полученных данных с помощью модели машинного обучения на базе нейронной сети, причем в ходе указанной обработки осуществляется:

преобразование вектора каждого токена, полученного на этапе b), в вектор вероятностей принадлежности соответствующего токена заданному классу;

преобразование вектора токена, полученного на этапе b), в вектор логарифмов вероятностей принадлежности соответствующего токена по меньшей мере одному подтипу класса из заданных подтипов класса;

объединение подряд идущих токенов с одинаковым классом по меньшей мере в одну именованную сущность;

определение принадлежности по меньшей мере одной именованной сущности по меньшей мере одному подтипу класса, причем определение выполняется с помощью этапа, при котором:

векторы логарифмов вероятностей принадлежности токенов по меньшей мере одному подтипу класса суммируются и сравниваются с заданным пороговым значением (при определении подтипа класса, имеющего множество подтипов); или

векторы логарифмов вероятностей принадлежности токенов по меньшей мере одному подтипу класса суммируются и сравниваются между собой, причем в ходе сравнения выбирается подтип, имеющий максимальную вероятность (при определении подтипа класса, имеющего только один подтип).

d) извлекают по меньшей мере одну именованную сущность с определенным по меньшей мере одним подтипом класса.

2. Способ по п.1, характеризующийся тем, что сегментация представляет собой разбиение текстовых данных по меньшей мере на одно предложение.

3. Способ по п.1, характеризующийся тем, что токенизация сегментированных входных текстовых данных содержит этап токенизации данных на подслова.

4. Способ по п.2, характеризующийся тем, что токенизация данных на подслова выполняется, по меньшей мере с использованием алгоритма токенизации, выбираемого из группы: алгоритм BPE, алгоритм WordPiece.

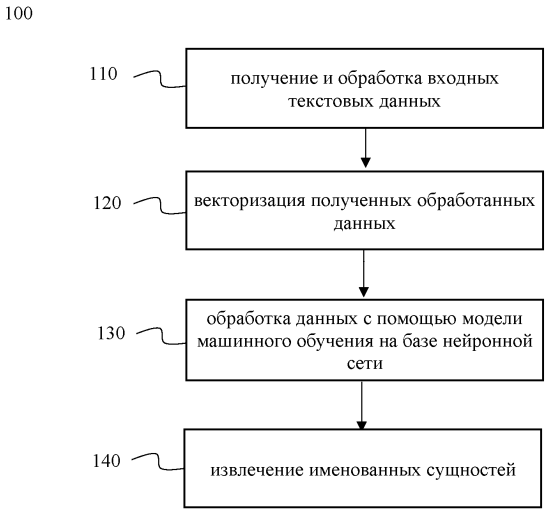
5. Способ по п.1, характеризующийся тем, что вектор каждого токена представляет собой вектор фиксированного размера.

6. Способ по п.1, характеризующийся тем, что дополнительно содержит этап отображения по меньшей мере одной извлеченной сущности.

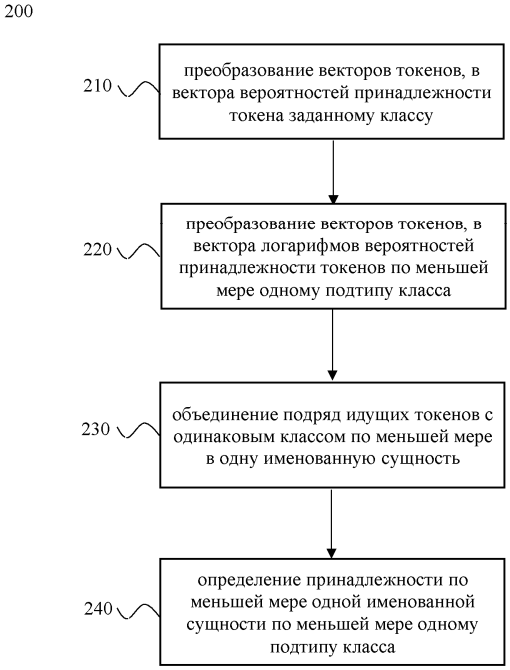
7. Система извлечения именованных сущностей из текстовой информации, содержащая:

по меньшей мере один процессор;

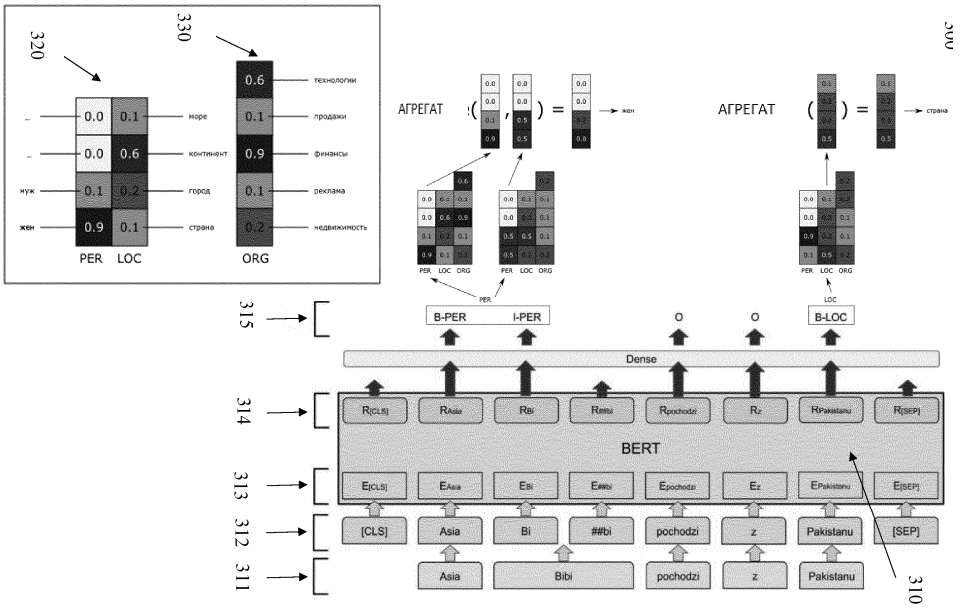
по меньшей мере одну память, соединенную с процессором, которая содержит машиночитаемые инструкции, которые при их выполнении по меньшей мере одним процессором обеспечивают выполнение способа по любому из пп.1-6.


100
110
получение и обработка входных текстовых данных
120
векторизация полученных обработанных данных
130
обработка данных с помощью модели машинного обучения на базе нейронной сети
140
извлечение именованных сущностей

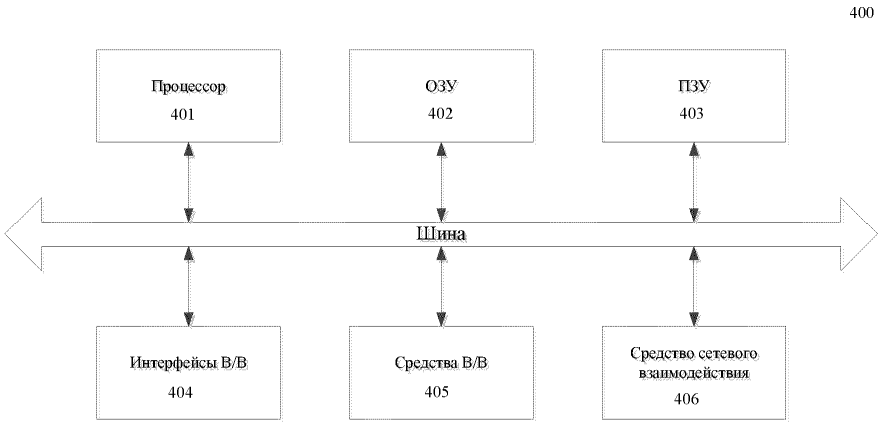
Фиг. 1


200
210
преобразование векторов токенов, в вектора вероятностей принадлежности токена заданному классу
220
преобразование векторов токенов, в вектора логарифмов вероятностей принадлежности токенов по меньшей мере одному подтипу класса
230
объединение подряд идущих токенов с одинаковым классом по меньшей мере в одну именованную сущность
240
определение принадлежности по меньшей мере одной именованной сущности по меньшей мере одному подтипу класса

Фиг. 2


320
330
PER
LOC
ORG
технологии
продажи
финансы
реклама
недвижимость
море
континент
город
страна
муж
жен
АГРЕГАТ
B-PER
I-PER
O
O
B-LOC
Dense
BERT
[CLS]
Asia
Bi
##bi
pochodzi
z
Pakistanu
[SEP]
Bibi
315
314
313
312
311
310

Фиг. 3


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

Фиг. 4