METHOD AND SYSTEM FOR TRAINING INFORMATION EXTRACTION MODEL AND INFORMATION EXTRACTION METHOD AND SYSTEM

TECHNICAL FIELD

[0001] This specification relates to the field of artificial intelligence technologies, and in particular, to a method and a system for training an information extraction model and an information extraction method and system.

BACKGROUND

[0002] Currently, information needs to be extracted from documents in many scenarios. For example, in a risk control scenario, to effectively prevent and control risky transactions, a risk control party requires transaction participants to upload proof documents that can prove authenticity of transactions. The proof documents include but are not limited to industrial and commercial proof documents, trade contracts, cross-border logistics proofs, and the like. The risk control party can perform transaction authenticity verification (that is, verify authenticity of transactions) based on the proof documents, to identify risk transactions. In the transaction authenticity verification process, information needs to be extracted from content in the proof documents.

[0003] Usually, an information extraction model can be obtained through pre-training, and information can be extracted from documents by using the information extraction model. However, a large quantity of labeled sample documents are required in a training process of the information extraction model. In actual application, sample documents described in a commonly used language (for example, Chinese and English) are relatively easy to obtain, while sample documents described in a less commonly used language are not easy to obtain. This makes a trained information extraction model applicable only to information extraction from documents in the commonly used language and not applicable to information extraction from documents in the less commonly used language.

[0004] Content in the background is merely information known to the inventor, and neither means that the information has entered the public domain before the application date of this disclosure, nor means that the information can become the existing technology of this disclosure.

SUMMARY

[0005] This specification provides a method and a system for training an information extraction model and an information extraction method and system, to obtain, through training based on a sample document described in a commonly used language, an information extraction model that supports information extraction from documents in the commonly used language and a less commonly used language.

[0006] To achieve the above objective, the following technical solutions are used in the embodiments of this specification:

[0007] According to a first aspect, this specification provides a method for training an information extraction model. The method can be used to train an information extraction model that has a multi-language information extraction capability. The method includes: obtaining raw text information corresponding to a sample document and information extraction labeling information corresponding to the sample document, where the sample document is a document described in at least one language in a first language set, the raw text information includes N raw text units in the sample document, and N is an integer greater than or equal to 1; generating augmented text information corresponding to the sample document based on the raw text information, where the augmented text information includes at least one augmented text unit corresponding to each of at least some of the N raw text units, and the augmented text unit is obtained by translating some text content in the raw text unit into at least one language in a second language set; and training the information extraction model based on the raw text information, the augmented text information, and the information extraction labeling information.

[0008] According to a second aspect, this specification further provides an information extraction method, including: obtaining a target document; performing text recognition on the target document, to obtain target text information, where the target text information includes at least one text unit in the target document; and extracting information from the target text information by using an information extraction model, to obtain key information in the target

document, where the information extraction model is obtained through training by using the method according to any implementation of the first aspect.

[0009] According to a third aspect, this specification further provides a system for training an information extraction model, including: at least one storage medium that stores at least one instruction set for training an information extraction model; and at least one processor that is communicatively connected to the at least one storage medium. When the training system runs, the at least one processor reads the at least one instruction set, and performs the following operations based on an indication of the at least one instruction set: obtaining raw text information corresponding to a sample document and information extraction labeling information corresponding to the sample document, where the sample document is a document described in at least one language in a first language set, the raw text information includes N raw text units in the sample document, and N is an integer greater than or equal to 1; generating augmented text information corresponding to the sample document based on the raw text information, where the augmented text information includes at least one augmented text unit corresponding to each of at least some of the N raw text units, and the augmented text unit is obtained by translating some text content in the raw text unit into at least one language in a second language set; and training the information extraction model based on the raw text information, the augmented text information, and the information extraction labeling information.

[0010] According to a fourth aspect, this specification further provides an information extraction system, including: at least one storage medium that stores at least one instruction set for information extraction; and at least one processor that is communicatively connected to the at least one storage medium. When the information extraction system runs, the at least one processor reads the at least one instruction set and implements the information extraction method provided in the second aspect.

[0011] According to a fifth aspect, this specification further provides a computer-readable nonvolatile storage medium. The computer-readable nonvolatile storage medium stores at least one instruction set, and when the at least one instruction set is executed by at least one processor, the method for training an information extraction model provided in the first aspect is implemented, or the information extraction method according to the second aspect is implemented.

[0012] Other functions of the method and the system for training an information extraction model and the information extraction method and system provided in this specification are

partially listed in the following descriptions. Creative aspects of the method and the system for training an information extraction model and the information extraction method and system provided in this specification can be fully explained by practice or using the methods, apparatuses, and combinations described in the following detailed examples.

BRIEF DESCRIPTION OF DRAWINGS

[0013] To describe the technical solutions in the embodiments of this specification more clearly, the following briefly describes the accompanying drawings needed for describing the embodiments. Clearly, the accompanying drawings in the following description show merely some embodiments of this specification, and a person of ordinary skill in the art can still derive other drawings from these accompanying drawings without creative efforts.

[0014] FIG. 1 is a schematic diagram of an application scenario of information extraction according to an embodiment of this specification;

[0015] FIG. 2 is a diagram of a hardware structure of a computing device according to an embodiment of this specification;

[0016] FIG. 3 is a flowchart of a method for training an information extraction model according to an embodiment of this specification;

[0017] FIG. 4 is a schematic diagram of a process of generating an augmented text unit according to an embodiment of this specification;

[0018] FIG. 5A is a schematic diagram of a group of raw text units and augmented text units according to an embodiment of this specification;

[0019] FIG. 5B is a schematic diagram of another group of raw text units and augmented text units according to an embodiment of this specification;

[0020] FIG. 5C is a schematic diagram of another group of raw text units and augmented text units according to an embodiment of this specification;

[0021] FIG. 6 is a schematic diagram of two training phases of an information extraction model according to an embodiment of this specification; and

[0022] FIG. 7 is a flowchart of an information extraction method according to an embodiment of this specification.

DESCRIPTION OF EMBODIMENTS

[0023] Specific application scenarios and requirements of this specification are provided in the following descriptions, so that a person skilled in the art can manufacture and use content of this specification. Various local modifications to the disclosed embodiments are clear to a person skilled in the art. In addition, the general principles defined herein can be applied to other embodiments and applications without departing from the spirit and scope of this specification. Therefore, this specification is not limited to the shown embodiments, but has a widest scope consistent with that of the claims.

[0024] The terms used herein are merely intended to describe specific example embodiments, and impose no limitation. For example, unless otherwise explicitly stated in the context, the singular forms "one", "an", and "the" used herein can include plural forms. When being used in this specification, the terms "include", "comprise", and/or "have" mean existence of an associated integer, step, operation, element, and/or component, but does not preclude existence of one or more other features, integers, steps, operations, elements, components, and/or groups or addition of other features, integers, steps, operations, elements, components, and/or groups to the system/method.

[0025] In consideration of the following descriptions, these and other features of this specification, operations and functions of related components of the structure, and economy of combination and manufacturing of components can be significantly improved. With reference to the accompanying drawings, all of these form a part of this specification. However, it should be clearly understood that the accompanying drawings are merely used for the purpose of illustration and description, and are not intended to limit the scope of this specification. It should be further understood that the accompanying drawings are not drawn to scale.

[0026] The flowcharts used in this specification show operations implemented by the system according to some embodiments of this specification. It should be clearly understood that the operations in the flowcharts may not be sequentially implemented. On the contrary, the operations can be implemented in a reverse sequence or simultaneously. In addition, one or more other operations can be added to the flowchart, and one or more operations can be removed from the flowchart.

[0027] For ease of description, the terms that appear below in this specification are first explained.

[0028] LayoutXLM: LayoutXLM is a pre-trained model. This model combines text and layout information to process cross-modal (for example, text and images) and cross-language tasks. LayoutXLM performs excellently in various cross-modal and cross-language benchmark tests, including tasks such as document image classification, table recognition, and table entity recognition.

[0029] Multi-language alignment task: The multi-language alignment task is usually a task of aligning or matching text or information in different languages. For example, in machine translation, sentences or phrases in source and target languages need to be aligned, so that a machine can learn of a correspondence between different languages. In cross-language information retrieval, queries and documents need to be aligned between different languages to find related information.

[0030] An application scenario of this specification is described below.

[0031] This specification provides a method for training an information extraction model and an information extraction method. The method for training an information extraction model and the information extraction method can be applied to a transaction authenticity verification scenario. In this scenario, a risk control party can obtain an information extraction model through pre-training. When a transaction participant initiates a transaction request, the risk control party can require the transaction participant to upload a proof document that can prove authenticity of the transaction. The risk control party extracts information from the proof document by using the information extraction model obtained through training, to obtain key information, and verifies the authenticity of the transaction based on the key information. If an authenticity verification result indicates that the transaction is a risky transaction, the risk control party intercepts the transaction. The method for training an information extraction model provided in this specification can be used in the training process of the information extraction model, and the information extraction method provided in this specification can be used in the process of extracting information from the proof document.

[0032] It should be noted that the transaction authenticity verification scenario is merely one of a plurality of use scenarios provided in this specification. The method for training an information extraction model and the information extraction method provided in this specification can not only be applied to the transaction authenticity verification scenario, but also be applied to all scenarios in which information needs to be extracted from a document, for example, a public

opinion analysis scenario, an intelligent paper marking scenario, a medical record analysis scenario, and a translation scenario. A person skilled in the art should understand that when the method for training an information extraction model and the information extraction method provided in this specification are applied to another use scenario, implementations and technical effects thereof are similar.

[0033] FIG. 1 is a schematic diagram of an application scenario of information extraction according to an embodiment of this specification.

[0034] As shown in FIG. 1, the application scenario 100 can include a training device 11 and an execution device 12. The application scenario 100 involves a model training phase and a model application phase. In the model training phase, the training device 11 can obtain an information extraction model through training based on a sample document library. The trained information extraction model can be deployed in the execution device 12, so that the execution device 12 can extract information from a target document by using the information extraction model.

[0035] The sample document library can include a sample document described in a commonly used language and information extraction labeling information corresponding to each sample document. The sample document library can be a database that is locally set, or can be a database that is set in a cloud or a server or at another location.

[0036] The training device 11 can be an electronic device that has a specific computing capability. The training device 11 can perform the method for training an information extraction model provided in this specification. The training device 11 can store data or instructions for performing the method for training an information extraction model described in this specification, and can execute or be configured to execute the data or the instructions. The training device 11 can include a hardware device having a data information processing function and a program required for driving the hardware device to work.

[0037] The information extraction model is a machine learning model that is obtained through training by the training device 11 and that has an information extraction capability. A specific structure of the information extraction model is not limited in this specification. In some embodiments, the information extraction model can be a LayoutXLM model.

[0038] In an example, the training device 11 can be a server, for example, a local server, a cloud server, a rack server, or a blade server; or can be a desktop computer, a notebook computer, or the like.

[0039] The execution device 12 can be an electronic device that has a specific computing capability. The execution device 12 can perform the information extraction method provided in this specification. The execution device 12 can store data or instructions for performing the information extraction method described in this specification, and can execute or be configured to execute the data or the instructions. The execution device 12 can include a hardware device having a data information processing function and a program required for driving the hardware device to work.

[0040] A trained information extraction model is deployed in the execution device 12. After obtaining a target document, the execution device 12 can extract information from the target document by using the information extraction model, to obtain key information in the target document.

[0041] In a system architecture of a specific scenario, the execution device 12 can be a specific device in the system architecture. A transaction authenticity verification scenario is used as an example. A system architecture of the scenario usually includes a terminal device (corresponding to a transaction client) and a server (corresponding to a transaction server).

[0042] In some embodiments, the execution device 12 can be a server. In this case, a transaction participant (for example, a user or a merchant) initiates a transaction request by using a terminal device, and uploads a transaction proof document by using the terminal device. The terminal device can send the transaction proof document to the server, and the server extracts information from the transaction proof document, to obtain key information, and then performs transaction authenticity verification based on the key information.

[0043] In some embodiments, the execution device 12 can alternatively be a terminal device. In this case, a transaction participant initiates a transaction request by using the terminal device, and uploads a transaction proof document by using the terminal device. The terminal device extracts information from the transaction proof document, to obtain key information, and sends the key information to a server, and the server performs transaction authenticity verification based on the key information.

[0044] In some embodiments, the terminal device can include a mobile device, a tablet computer, a notebook computer, a built-in device in a motor vehicle or similar content, or any combination thereof. In some embodiments, the mobile device can include a smart home device, a smart mobile device, a virtual reality device, an augmented reality device or a similar device, or

any combination thereof. In some embodiments, the smart home apparatus can include a smart television, a desktop computer, or any combination. In some embodiments, the smart mobile device can include a smartphone, a personal digital assistant, a game device, a navigation device, or any combination thereof. In some embodiments, the virtual reality device or the augmented reality device may include a virtual reality helmet, virtual reality glasses, a virtual reality patch, an augmented reality helmet, augmented reality glasses, an augmented reality patch or similar content, or any combination thereof. For example, the virtual reality device or the augmented reality device may include Google glasses, a head mounted display, or VR. In some embodiments, the built-in apparatus in the motor vehicle can include a vehicle-mounted computer, a vehicle-mounted television, or the like. In some embodiments, the terminal device can include an image capture device, configured to capture video or image information of a target document. In some embodiments, the image capture device can be a two-dimensional image capture device (for example, an RGB camera), or can be a two-dimensional image capture device (for example, an RGB camera) and a depth image capture device (for example, a 3D structured light camera or a laser detector). In some embodiments, the image capture device can be built into the terminal device, for example, a camera built into a smartphone and a tablet computer. In some embodiments, the image capture device can alternatively be an external device connected to the terminal device, for example, a scanner, a high-speed photographic apparatus, or a camera connected to a desktop computer.

[0045] In some embodiments, one or more application programs (apps) can be installed in the terminal device. The app includes but is not limited to a web browser app program, a search app program, a chat app program, a shopping app program, a video app program, a financial app program, an instant messaging tool, a mailbox client, social platform software, and the like. In some embodiments, a target app can be installed in the terminal device. The target app can perform text recognition on the target document based on the video or image information of the target document captured by the terminal device, to obtain target text information. Then, the terminal device inputs the target text information to the information extraction model, to obtain the key information in the target document.

[0046] It should be noted that the training device 11 and the execution device 12 can correspond to a same physical device, or can correspond to different physical devices. This is not limited in this specification.

[0047] FIG. 2 is a diagram of a hardware structure of a computing device according to an embodiment of this specification.

[0048] In some embodiments, the computing device 200 can serve as the training device 11 in FIG. 1 to perform the method for training an information extraction model provided in this specification, and the computing device 200 can further serve as the system for training an information extraction model provided in this specification. In some embodiments, the computing device 200 can serve as the execution device 12 in FIG. 1 to perform the information extraction method provided in this specification, and the computing device 200 can further serve as the information extraction system provided in this specification.

[0049] As shown in FIG. 2, the computing device 200 can include at least one storage medium 230 and at least one processor 220. In some embodiments, the computing device 200 can further include a communication port 250 and an internal communication bus 210. The computing device 200 can further include an I/O component 260.

[0050] The internal communication bus 210 can be connected to different system components. For example, the internal communication bus 210 can be connected to the storage medium 230, the processor 220, the communication port 250, the I/O component 260, and the like.

[0051] The I/O component 260 supports input/output between the computing device 200 and other components.

[0052] The communication port 250 is used for data communication between the computing device 200 and the outside. For example, the communication port 250 can be used for data communication between the computing device 200 and a network 13. The communication port 250 can be a wired communication port, or can be a wireless communication port.

[0053] The storage medium 230 can include a data storage apparatus. The data storage apparatus can be a non-transitory storage medium, or can be a transitory storage medium. For example, the data storage apparatus can include one or more of a disk 232, a read-only storage medium (ROM) 234, or a random access storage medium (RAM) 236. The storage medium 230 further includes at least one instruction set stored in the data storage apparatus. The instruction set can include computer program code. The computer program code can include a program, a routine, an object, a component, a data structure, a process, a module, and the like.

[0054] The at least one processor 220 can be communicatively connected to the at least one storage medium 230. When the computing device 200 runs, the at least one processor 220 reads

the at least one instruction set, and performs, based on an indication of the at least one instruction set, the method for training an information extraction model or the information extraction method provided in this specification. The processor 220 can perform the steps included in the method for training an information extraction model or the information extraction method. The processor 220 can be in a form of one or more processors. In some embodiments, the processor 220 can include one or more hardware processors, for example, a microcontroller, a microprocessor, a reduced instruction set computer (RISC), an application-specific integrated circuit (ASIC), an application-specific instruction set processor (ASIP), a central processing unit (CPU), a graphics processing unit (GPU), a physical processing unit (PPU), a microcontroller unit, a digital signal processor (DSP), a field programmable gate array (FPGA), an advanced RISC machine (ARM), a programmable logic device (PLD), any circuit, processor, or the like that can perform one or more functions, or any combination thereof.

[0055] To merely describe a problem, only one processor 220 is shown in the computing device 200 in the accompanying drawings. However, it should be noted that the computing device 200 in this specification can further include a plurality of processors. Therefore, the operations and/or method steps disclosed in this specification can be performed by one processor, or can be jointly performed by a plurality of processors. For example, if the processor 220 in the computing device 200 described in this specification performs step A and step B, it should be understood as that step A and step B can be jointly or separately performed by two different processors 220 (for example, a first processor performs step A, and a second processor performs step B, or a first processor and a second processor jointly perform step A and step B).

[0056] FIG. 3 is a flowchart of a method for training an information extraction model according to an embodiment of this specification. As described above, the training device 11 can perform the method for training an information extraction model in this specification.

[0057] As shown in FIG. 3, the method for training an information extraction model can include the following steps.

[0058] S310: Obtain raw text information corresponding to a sample document and information extraction labeling information corresponding to the sample document, where the sample document is a document described in at least one language in a first language set, the raw text information includes N raw text units in the sample document, and N is an integer greater than or equal to 1.

[0059] The first language set can be a set including at least one commonly used language. In this specification, a commonly used language can be understood as a frequently used language, and a less commonly used language can be understood as a less frequently used language. It can be understood that different regions or environments may correspond to different commonly used languages. In some embodiments, the first language set can include at least one of Chinese and English.

[0060] When the first language set includes Chinese and English, the sample document is a document described in at least one language in Chinese and English. A document type of the sample document is not limited in this specification. For example, the sample document can be a contract document, a personal identity document, a company or organization license, a permit, a logistics document, a flow document, or the like. The sample document can be a physical document, an electronic version or a photo of a physical document, or the like.

[0061] The raw text information corresponding to the sample document is text information included in the sample document. The raw text information can include a plurality of raw text units. In this specification, one raw text unit refers to one text block that can express a specific semantic meaning. One raw text unit can correspond to one character/word, phrase, sentence, paragraph, or the like in the sample document. A quantity of raw text units included in the raw text information is not limited in this specification. The N raw text units are used below as an example for description.

[0062] In some embodiments, text recognition can be performed on the sample document to obtain the N raw text units. For example, optical character recognition (OCR) can be performed on the sample document to obtain the N raw text units. For example, a character/word, a sentence, or a paragraph recognized from the sample document through OCR can be used as a raw text unit. In some embodiments, layout information of the N raw text units can be further obtained by performing text recognition on the sample document. The layout information can represent a location, a direction, a format, and the like of the raw text unit in the sample document. That is, the raw text information can include the N raw text units and layout information corresponding to each of the N raw text units.

[0063] In some embodiments, when an English text "Account Name: HONGKONG ABC TECHNOLOGY CO., LIMITED" is recognized from the sample document through OCR, the English text can be used as a raw text unit.

[0064] In some other embodiments, each phrase in the above English text can be used as a raw text unit. For example, the above English text includes three raw text units "Account Name", "HONGKONG ABC TECHNOLOGY", and "CO., LIMITED".

[0065] It should be noted that the above example of the raw text unit is merely an example, and a finer granularity can be further set. For example, in some embodiments, "HONGKONG ABC TECHNOLOGY" can be further segmented to obtain three raw text units "HONGKONG", "ABC", and "TECHNOLOGY". In this way, the above English text includes a total of five raw text units. Alternatively, in some other embodiments, a coarser granularity can be set. For example, the above English text includes a total of two raw text units: "Account Name" and "HONGKONG ABC TECHNOLOGY CO., LIMITED".

[0066] In some embodiments, the information extraction labeling information corresponding to the sample document is used to indicate key information in the sample document.

[0067] The key information is information to which attention needs to be paid in the sample document, for example, information that needs to be extracted by the information extraction model. In an example, a transaction authenticity verification scenario is used as an example. The key information can be set based on a type of a document. For example, for a personal identity document, the key information can include a name, a document number, a document validity period, and the like. For a logistics document, the key information can include a logistics document number, a place of shipment, a place of receipt, a shipper, a consignee, and the like.

[0068] In an example, to indicate the key information in the sample document, the information extraction labeling information can include an identifier of a raw text unit corresponding to the key information in the sample document. Alternatively, the information extraction labeling information can include text content in a raw text unit corresponding to the key information in the sample document.

[0069] In an example, when the sample document is a logistics document, OCR recognition can be performed on an image of the logistics document. The following raw text units may be obtained: "1: Logistics document number: 123456789", "2: Place of shipment: City B, Country A", "3: Place of receipt: City D, Country C", "4: Shipper: Company E", "5: Consignee: Company F", "6: Shipment weight: xxx KG", "7: Shipment size: x*y*z", "8: Transportation mode: Express delivery", "9: Expense details: Unit price: a yuan, freight: b yuan, and insurance: c yuan", and the like.

[0070] The key information can include the following five raw text units: "1: Logistics document number: 123456789", "2: Place of shipment: City B, Country A", "3: Place of receipt: City D, Country C", "4: Shipper: Company E", and "5: Consignee: Company F". In this case, identifiers (that is, numbers) "1, 2, 3, 4, and 5" of the five raw text units can be used as the information extraction labeling information to indicate that the five raw text units are the key information.

[0071] The information extraction labeling information can be obtained through manual labeling, or can be automatically labeled by using a semantic recognition algorithm. This is not limited in this specification.

[0072] In some embodiments, a plurality of sample documents can be collected in advance, text recognition can be performed on the plurality of sample documents to obtain raw text information corresponding to each sample document, and key information labeling can be performed on the plurality of sample documents to obtain information extraction labeling information. The raw text information and the information extraction labeling information corresponding to each sample document can be stored in a sample document library. Therefore, when the training device 11 needs to train the information extraction model, the raw text information and the information extraction labeling information corresponding to the sample document are obtained from the sample document library.

[0073] S320: Generate augmented text information corresponding to the sample document based on the raw text information, where the augmented text information includes at least one augmented text unit corresponding to each of at least some of the N raw text units, and the augmented text unit is obtained by translating some text content in the raw text unit into at least one language in a second language set.

[0074] The second language set can be a set including at least one less commonly used language. In some embodiments, a language in the second language set is different from a language in the first language set. For example, if the first language set includes Chinese and English, the second language set can include Japanese, Korean, French, and Russian. In some embodiments, the second language set includes a language other than that in the first language set in all languages. For example, if the first language set includes Chinese and English, the second language set includes a language other than Chinese and English in all the languages.

[0075] In this specification, the training device can translate at least some raw text units in the

raw text information (translate the raw text units into a language in the second language set), to obtain the augmented text information. The translated text unit is referred to as an augmented text unit. For a meaning of the augmented text unit, refer to the above explanation of the raw text unit. Details are not described herein. A difference between the augmented text unit and the raw text unit is that the raw text unit is a raw text unit recognized from the sample document, and the augmented text unit is a text unit obtained after the raw text unit is translated.

[0076] In some embodiments, the augmented text information includes at least one augmented text unit corresponding to each of M of the N raw text units, and M is an integer greater than or equal to 1. That is, the augmented text information is obtained by translating the M raw text units in the raw text information. A percentage of the M raw text units in the N raw text units is less than or equal to a preset threshold.

[0077] For example, for a sample document, if the preset threshold is n%, M can be an integer less than or equal to N*n%.

[0078] For example, n% can be 20%–30%. In an example of 20%, when N is 300, M is less than or equal to 60. That is, when the sample document includes 300 raw text units, a maximum of 60 raw text units can be selected for translation, to obtain corresponding augmented text units. Raw text units with a percentage less than or equal to the preset threshold are translated to obtain corresponding augmented text units, so that it can be ensured that the augmented text unit obtained through translation can still express a meaning of the raw text unit. In addition, the augmented text unit and the raw text unit can be semantically aligned across different language sets. This helps improve performance of an information extraction model obtained through pre-training based on the augmented text unit, so that the information extraction model obtained through pre-training has a better capability of representing documents in a plurality of languages.

[0079] With reference to FIG. 4, a manner of generating the augmented text unit is described below by using an example.

[0080] FIG. 4 is a schematic diagram of a process of generating an augmented text unit according to an embodiment of this specification. As shown in FIG. 4, in some embodiments, the training device can perform word segmentation processing on the raw text unit, to obtain a plurality of text subunits. For example, with reference to FIG. 4, it is assumed that the raw text unit is divided into a text subunit 1, a text subunit 2, ..., and a text subunit X. The training device can determine at least one target text subunit from the plurality of text subunits. For example, the

training device can select the text subunit 2 and the text subunit 4 as target text subunits. Further, the training device can translate the at least one target text subunit into at least one language in the second language set, to obtain the augmented text unit. For example, with reference to FIG. 4, the training device translates the text subunit 2 to obtain a translation corresponding to the text subunit 2, and translates the text subunit 4 to obtain a translation corresponding to the text subunit 4. Further, the training device generates the augmented text unit based on the text subunit 1, the translation corresponding to the text subunit 2, the text subunit 3, the translation corresponding to the text subunit 4, the text subunit 5, ..., and the text subunit X.

[0081] By translating some text content in the raw text unit from a language in the first language set into a language in the second language set, augmented text information in more languages can be generated based on the existing raw text information, so that the sample document corresponds to text information in a plurality of languages, to resolve a problem that a sample document in a less commonly used language is not easy to obtain. Further, because the augmented text unit is obtained by translating some text content in the raw text unit, the raw text unit and each text subunit in the augmented text unit are semantically aligned. This helps the information extraction model execute a multi-language alignment task in a pre-training process described below, so that a pre-trained information extraction model has a capability of representing a plurality of languages.

[0082] In some embodiments, when performing word segmentation processing on the raw text unit, the training device can perform word segmentation based on a rule, perform word segmentation based on statistics, or perform word segmentation based on deep learning. For example, for English, word segmentation can be performed based on an English syntax rule. For example, word segmentation is performed by using a space and a punctuation mark in an English text, to obtain a word segmentation result (that is, a plurality of text subunits). For another example, for Chinese, statistics on a large amount of text data can be collected by using a word segmentation algorithm, to learn a word boundary, and then a Chinese text can be input to the word segmentation algorithm, to obtain a word segmentation result. The word segmentation algorithm can include a hidden Markov model (HMM), a conditional random field (CRF), and the like.

[0083] In some embodiments, when performing word segmentation processing on the raw text unit, the training device can implement word segmentation by training a neural network-based model. For example, a sequence labeling model based on a bidirectional long short-term memory

(BiLSTM) can be used for training, and then the raw text unit can be input to the BiLSTM, to obtain a word segmentation result.

[0084] For example, after word segmentation is performed on the raw text unit "Account Name: HONGKONG ABC TECHNOLOGY CO., LIMITED", the following word segments can be obtained: "Account Name", "HONGKONG", "ABC", "TECHNOLOGY", and "CO., LIMITED".

[0085] Each word segment obtained after word segmentation processing is one text subunit of the raw text unit.

[0086] In some embodiments, when determining the target text subunit, the training device can determine the at least one target text subunit from the plurality of text subunits based on at least one of occurrence frequencies of the plurality of text subunits in the sample document and parts of speech of the plurality of text subunits.

[0087] In an example, the training device can determine a verb and/or a noun with a quantity of occurrence times greater than or equal to a preset threshold in the plurality of text subunits as the target text subunit.

[0088] With reference to the above example, there are nouns "Account Name" and "LIMITED" in the word segments. When quantities of occurrence times of "Account Name" and "CO., LIMITED" in the sample document are greater than the preset threshold, it can be determined that "Account Name" and "CO., LIMITED" are target text subunits.

[0089] For example, if the preset threshold is 10, when the quantity of occurrence times of "Account Name" is greater than 10 and the quantity of occurrence times of "CO., LIMITED" is greater than 10, it can be determined that "Account Name" and "CO., LIMITED" are target text subunits.

[0090] In another example, the training device can rank the plurality of text subunits based on a word frequency, and then determine a verb and/or a noun ranked before a preset ranking as the target text subunit.

[0091] With reference to the above example, there are nouns "Account Name" and "CO., LIMITED" in the word segments. After the text subunits in the sample document are ranked based on the word frequency, if "Account Name" and "CO., LIMITED" are ranked before the preset ranking, it can be determined that "Account Name" and "LIMITED" are target text subunits.

[0092] For example, if the preset ranking is 10, when "Account Name" is ranked third and "CO., LIMITED" is ranked fifth, it can be determined that "Account Name" and "CO., LIMITED"

are target text subunits.

[0093] In some embodiments, the training device can directly randomly determine at least one text subunit as the target text subunit from the plurality of text subunits.

[0094] With reference to the above example, at least one text subunit can be randomly determined as the target text subunit from "Account Name", "HONGKONG", "ABC", "TECHNOLOGY", and "CO., LIMITED".

[0095] For example, a word list can be generated based on the plurality of text subunits, and each text subunit is one word in the list. The list is numbered. For example, a number 0 corresponds to "Account Name", a number 1 corresponds to "HONGKONG", a number 2 corresponds to "ABC", a number 3 corresponds to "TECHNOLOGY", and a number 4 corresponds to "CO., LIMITED". Then, one to five non-repeated random numbers from 0 to 4 are generated, and a text subunit corresponding to the random number is used as the target text subunit. If two random numbers "0" and "4" are generated, it is determined that "Account Name" and "CO., LIMITED" are target text subunits.

[0096] Compared with words of other parts of speech, nouns and/or verbs are usually semantically significant in documents and are less prone to ambiguity in translation. Compared with words with a relatively low occurrence frequency, words with a relatively high occurrence frequency are semantically significant in documents. The target text subunit is determined based on at least one of the occurrence frequencies of the text subunits in the sample document and the parts of speech of the plurality of text subunits, so that a representative and semantically significant word can be selected and a word that is not semantically significant can be blocked, to ensure accuracy during subsequent translation and improve accuracy of an information extraction model obtained through training.

[0097] In some embodiments, after determining the target text subunit from the raw text unit, the training device can translate the at least one target text subunit into at least one language in the second language set, to obtain the augmented text unit.

[0098] In an example, the second language set can include a plurality of languages. When the target text subunit is translated, at least one language in the second language set can be randomly selected. For example, with reference to the above example of the second language set, the target text subunit can be translated into at least one of Japanese, Korean, French, and Russian.

[0099] FIG. 5A is a schematic diagram of a group of raw text units and augmented text units

according to an embodiment of this specification. As shown in FIG. 5A, in an example, if the raw text unit is "Account Name: HONGKONG ABC TECHNOLOGY CO., LIMITED", the determined target text subunits are "Account Name" and "CO., LIMITED", and the translation language determined from the second language set is Japanese, the obtained augmented text unit is "アカウント名: HONGKONG ABC TECHNOLOGY 有限会社".

[0100] FIG. 5B and FIG. 5C are schematic diagrams of another group of raw text units and augmented text units according to an embodiment of this specification. As shown in FIG. 5B and FIG. 5C, if the raw text unit is "Account Name: HONGKONG ABC TECHNOLOGY CO., LIMITED", the determined target text subunits are "Account Name" and "CO., LIMITED", and the translation language determined from the second language set is Japanese and Korean, the augmented text unit obtained based on the above example can be "アカウント名: HONGKONG ABC TECHNOLOGY 유한회사" shown in FIG. 5B, or can be "계정 이름: HONGKONG ABC TECHNOLOGY 有限会社" shown in FIG. 5C.

[0101] S330: Train the information extraction model based on the raw text information, the augmented text information, and the information extraction labeling information.

[0102] In this specification, a training process of the information extraction model can include two phases: pre-training and fine-tuning training. In each phase, the information extraction model can be indicated to execute a different task, to train a different capability of the information extraction model. For example, in the pre-training phase, the information extraction model can be indicated, by using a header file before outputting, to execute a multi-language alignment task, and in the fine-tuning training phase, the information extraction model can be indicated, by using a header file before outputting, to execute an information extraction task.

[0103] The two training phases are described below with reference to FIG. 6. FIG. 6 is a schematic diagram of two training phases of an information extraction model according to an embodiment of this specification.

[0104] With reference to FIG. 6, in the pre-training phase, the training device can pre-train the information extraction model based on the raw text information and the augmented text information. A multi-language alignment task can be executed in a pre-training process, so that a pre-trained information extraction model has a capability of representing a target text unit, where the target text unit is a text unit described in any language in the first language set and the second language set.

[0105] In an example, the information extraction model can be LayoutXLM. During pre-training, the training device can input the raw text information and the augmented text information to LayoutXLM. Then, the training device executes the multi-language alignment task for the LayoutXLM model and performs pre-training.

[0106] In some embodiments, the training device can obtain a plurality of sample pairs based on the raw text information and the augmented text information, where one sample pair includes one raw text unit and one augmented text unit that have a correspondence. Then, the training device can pre-train the information extraction model through contrastive learning based on the plurality of sample pairs.

[0107] A raw text unit and an augmented text unit in a sample pair can be considered as being of a same type, and the sample pair can be considered as being of a different type for another sample pair. For example, it is assumed that the plurality of sample pairs include "A1, A2", "B1, B2", "C1, C2", and "D1, D2". A1 and A2 are a corresponding raw text unit (A1) and a corresponding augmented text unit (A2), A1 and A2 are of a same type, A1 and B1, B2, C1, C2, D1, and D2 are of different types, and A2 and B1, B2, C1, C2, D1, and D2 are of different types. Correspondingly, B1 and B2 are of a same type, B1 and A1, A2, C1, C2, D1, and D2 are of different types, and B2 and A1, A2, C1, C2, D1, and D2 are of different types.

[0108] In some embodiments, the training device can input the plurality of sample pairs to the information extraction model, to obtain a plurality of feature vector groups, where each feature vector group includes feature vectors respectively corresponding to two text units in a sample pair.

[0109] For example, the training device inputs the plurality of sample pairs to the information extraction model, and the information extraction model can extract feature vectors of the plurality of sample pairs, to obtain a plurality of feature vector groups. For example, after the information extraction model performs feature extraction on the sample pair "A1, A2", a feature vector group "a1, a2" can be obtained, where a1 is a feature vector of the raw text unit (A1), and a2 is a feature vector of the augmented text unit (A2). Similarly, for the plurality of sample pairs, the following feature vector groups can be extracted: "b1, b2", "c1, c2", and "d1, d2".

[0110] In some embodiments, after obtaining the feature vector group corresponding to the sample pair, the training device can update a parameter of the information extraction model based on a contrastive learning strategy. A contrastive learning process can include: updating the parameter of the information extraction model with a training objective of minimizing a difference

between two feature vectors in a same feature vector group and maximizing a difference between feature vectors in different feature vector groups.

[0111] In some embodiments, in a vector space, a difference between two vectors can be represented by using a distance. A shorter distance represents a smaller difference, and a longer distance represents a larger difference.

[0112] In an example, a distance in the vector space can be measured by using a norm, and common norms include an L1 norm, an L2 norm, and an infinite norm. For example, the L1 norm is also referred to as a Manhattan norm or a sum of absolute values, which is a sum of absolute values of elements in a vector. The L2 norm is also referred to as a Euclidean norm, which is a square root of a sum of squares of elements in a vector, that is, a Euclidean distance. The infinite norm is a maximum value in absolute values of elements in a vector. The L2 norm is used as an example. An L2 norm of two vectors is a Euclidean distance between the two vectors.

[0113] By minimizing a distance between two feature vectors in a same feature vector group and maximizing a distance between feature vectors in different feature vector groups, a distance between two text units in a same sample pair in the vector space can be shortened, and the two text units in the same sample pair are used to indicate representations of a same semantic meaning in different languages. In this way, a document representation capability learned by the information extraction model from a document in a commonly used language (the first language set) can be migrated to a less commonly used language (the second language set) to implement multi-language alignment, so as to improve a capability of representing documents in a plurality of languages.

[0114] With further reference to FIG. 6, after the pre-training phase, the training device can perform fine-tuning training on the pre-trained information extraction model based on the raw text information and the information extraction labeling information, so that an information extraction model obtained after fine-tuning training has a capability of extracting key information.

[0115] In an example, when performing fine-tuning training, the training device can input the raw text information to pre-trained LayoutXLM. Then, the information extraction task is executed by using the pre-trained LayoutXLM model, and fine-tuning training is performed.

[0116] In some embodiments, the training device can input the plurality of raw text units to the pre-trained information extraction model, to obtain feature vectors respectively corresponding to the plurality of raw text units by using the information extraction model, and perform key information extraction based on the feature vectors respectively corresponding to the plurality of

raw text units, to obtain an information extraction result. Further, the training device can update the parameter of the information extraction model based on a supervised learning strategy. For example, the training device updates the parameter of the information extraction model with an objective of minimizing a difference between the information extraction result and the information extraction labeling information.

[0117] In some embodiments, with reference to FIG. 6, when the information extraction model is the LayoutXLM model, the LayoutXLM model can support fusion of vision and language, that is, the LayoutXLM model can support a text as an input, and can further support layout information as an input. Therefore, the training device can input the plurality of raw text units and the layout information of the plurality of raw text units to the pre-trained information extraction model, to obtain feature vectors respectively corresponding to the plurality of pieces of raw text information by using the information extraction model, and further perform key information extraction based on the feature vectors respectively corresponding to the plurality of raw texts, to obtain an information extraction result. Further, the training device can update the parameter of the information extraction model based on a supervised learning strategy. In this training manner, in the fine-tuning training phase, the information extraction model can not only understand a document from a language dimension based on the raw text unit, but also understand the document from a visual dimension based on the layout information. This can further improve an information understanding capability of the information extraction model for the document, thereby improving information extraction performance.

[0118] In conclusion, in the method for training an information extraction model provided in this specification, the raw text information corresponding to the sample document described in the language in the first language set can be obtained; some text content in the raw text information can be translated into the language in the second language set, to obtain the augmented text information; and then the information extraction model can be trained based on the raw text information, the augmented text information, and the information extraction labeling information of the sample document. By translating the raw text information, the augmented text information that has a same meaning but includes a different language can be obtained, and the information extraction model is trained by using the raw text information and the augmented text information, so that the information extraction model obtained through training can have a capability of extracting information in a plurality of languages, that is, is not only applicable to information

extraction from a document described in a language (for example, a commonly used language) in the first language set, but also applicable to information extraction from a document described in a language (for example, a less commonly used language) in the second language set, to meet a requirement of extracting information from documents in a plurality of languages.

[0119] According to another aspect of this specification, an information extraction method is provided. FIG. 7 is a flowchart of an information extraction method according to an embodiment of this specification. In an example, the execution device 12 in FIG. 1 can perform the information extraction method in this specification.

[0120] As shown in FIG. 7, the information extraction method can include the following steps.

[0121] S710: Obtain a target document.

[0122] In some embodiments, the target document can be a file, a certificate, a document, or the like described in any one or more languages in a first language set and a second language set. The target document can be a physical document, an electronic version or a photo of a physical document, or the like.

[0123] For example, in a transaction authenticity verification scenario, the target document can be a transaction proof document that is uploaded by a transaction participant to prove authenticity of a target transaction. When the execution device 12 serves as a terminal device, the execution device 12 can obtain the target document based on the upload operation of the transaction participant. When the execution device 12 serves as a server, the execution device 12 can receive the target document from the terminal device.

[0124] S720: Perform text recognition on the target document, to obtain target text information, where the target text information includes at least one text unit in the target document.

[0125] For a process in which the execution device 12 performs text recognition on the target document, refer to the text recognition process of the sample document described above. Details are not described herein.

[0126] In some embodiments, the target text information can include layout information of each text unit in addition to the at least one text unit in the target document. The layout information can represent a location, a direction, a format, and the like of the text unit in the target document.

[0127] S730: Extract information from the target text information by using an information extraction model, to obtain key information in the target document.

[0128] The information extraction model is obtained through training by using the above

method for training an information extraction model. The information extraction model is deployed in the execution device 12. The execution device 12 can input the target text information to the information extraction model, to extract information by using the information extraction model, to obtain the key information in the target document.

[0129] In the transaction authenticity verification scenario, the key information in the target document can be used to verify the authenticity of the transaction.

[0130] For example, a risk control party can pre-configure, on the server, the information extraction model obtained through training by using the above method for training an information extraction model. When the transaction participant initiates a transaction request, the transaction participant can upload a proof document that can prove authenticity of the transaction, for example, a transaction contract and a logistics document.

[0131] The risk control party can obtain key information in the proof document by using the information extraction model, and verify the authenticity of the transaction based on the key information.

[0132] In an example, the transaction participant can obtain an electronic version of the proof document and upload the electronic version to the risk control party, and the risk control party inputs the received proof document to the information extraction model, and obtains the key information in the proof document.

[0133] For example, when the proof document is a logistics document, the obtained key information can include a shipper and a consignee. When the shipper in the logistics document is consistent with a receiver of the current transaction, and the consignee in the logistics document is consistent with a payer of the current transaction, it can be determined that the current transaction is a real transaction. Otherwise, it can be determined that the current transaction is not a real transaction. In some cases, the key information extracted from the logistics document can further include a logistics document number in addition to the shipper and the consignee. The risk control party can query, based on the logistics document number, whether a logistics status is a "delivered" state. When the logistics status is the "delivered" state, the shipper in the logistics document is consistent with the receiver of the current transaction, and the consignee in the logistics document is consistent with the payer of the current transaction, it can be determined that the current transaction is a real transaction. When any one of the above three conditions is not met, it can be determined that the current transaction is not a real transaction. Further, when determining that the

current transaction is a real transaction, the risk control party can release the current transaction. When determining that the current transaction is not a real transaction, the risk control party can intercept the current transaction.

[0134] It should be noted that the above manner of verifying the authenticity of the transaction based on the extracted key information is merely an example description. This is not limited in this specification. After the key information is extracted, transaction authenticity verification can alternatively be performed in another manner.

[0135] According to another aspect of this specification, a computer-readable non-transitory storage medium is provided and stores at least one instruction set for training an information extraction model or information extraction. When the at least one instruction set is executed by a processor, the at least one instruction set instructs the processor to implement the steps of the method for training an information extraction model or the information extraction method described in this specification. In some possible implementations, various aspects of this specification can be further implemented in a form of a program product, including program code. When the program product runs on a computing device 200, the program code is used to enable the computing device 200 to perform the steps of the method for training an information extraction model or the information extraction method described in this specification. The program product configured to implement the above method can use a portable compact disc read-only memory (CD-ROM) that includes program code, and can run on the computing device 200. However, the program product in this specification is not limited thereto. In this specification, a readable storage medium can be any tangible medium that includes or stores a program, and the program can be used by or in combination with an instruction execution system. The program product can be any combination of one or more readable media. The readable medium can be a readable signal medium or a readable storage medium. For example, the readable storage medium can be but is not limited to an electrical system, apparatus, or device, a magnetic system, apparatus, or device, an optical system, apparatus, or device, an electromagnetic system, apparatus, or device, an infrared system, apparatus, or device, or a semiconductor system, apparatus, or device, or any combination thereof. More specific examples of the readable storage medium include an electrical connection including one or more conducting wires, a portable disk, a hard disk, a random access memory (RAM), a read-only memory (ROM), an erasable programmable read-only memory (EPROM) or a flash memory, an optical fiber, a portable compact disc read-only memory (CD-

ROM), an optical storage device, a magnetic storage device, or any proper combination thereof. The computer-readable storage medium can include a data signal propagated in a baseband or as a part of a carrier, and carries readable program code. The propagated data signal can be in a plurality of forms, and includes but is not limited to an electromagnetic signal, an optical signal, or any proper combination thereof. The readable storage medium can alternatively be any readable medium other than the readable storage medium, and the readable medium can send, propagate, or transmit a program used by or in combination with an instruction execution system, apparatus, or device. The program code included in the readable storage medium can be transmitted by using any proper medium, including but not limited to a wireless medium, a wired medium, an optical cable, an RF medium, or any proper combination thereof. Program code for performing the operations in this specification can be written by using any combination of one or more program design languages. The program design language includes an object-oriented program design language, for example, Java or C++, and further includes a conventional procedural program design language, for example, "C" language, or a similar program design language. The program code can be completely executed on the computing device 200, partially executed on the computing device 200, executed as an independent software package, partially executed on the computing device 200 and partially executed on a remote computing device, or completely executed on a remote computing device.

[0136] Specific embodiments of this specification are described above. Other embodiments fall within the scope of the appended claims. In some cases, the actions or steps described in the claims can be performed in a sequence different from that in the embodiments and desired results can still be achieved. In addition, the process depicted in the accompanying drawings does not necessarily need a particular sequence or consecutive sequence to achieve the desired results. In some implementations, multitasking and parallel processing are possible or may be advantageous.

[0137] In conclusion, after reading the detailed disclosure, a person skilled in the art can understand that the detailed disclosure can be presented only by using an example, and can impose no limitation. Although it is not explicitly stated herein, a person skilled in the art can understand that the requirements of this specification include various proper changes, improvements, and modifications to the embodiments. These changes, improvements, and modifications are intended to be provided in this specification, and fall within the spirit and scope of the example embodiments of this specification.

[0138] In addition, some terms in this specification are used to describe the embodiments of this specification. For example, "one embodiment", "embodiments", and/or "some embodiments" mean/means that specific features, structures, or characteristics described with reference to the embodiment can be included in at least one embodiment of this specification. Therefore, it can be emphasized and understood that two or more of "embodiments", "one embodiment", and "alternative embodiments" in parts of this specification do not necessarily indicate a same embodiment. In addition, specific features, structures, or characteristics can be properly combined in one or more embodiments of this specification.

[0139] It should be understood that in the above descriptions of the embodiments of this specification, to help understand a feature, for the purpose of simplifying this specification, various features are combined in a single embodiment, accompanying drawings, or descriptions thereof in this specification. However, this does not mean that the combination of these features is necessary. It is entirely possible for a person skilled in the art to label some devices and understand the devices as separate embodiments when reading this specification. That is, the embodiments of this specification can also be understood as an integration of a plurality of secondary embodiments. Content of each secondary embodiment is also true when a quantity of features is less than a quantity of all features in one of the above disclosed embodiments.

[0140] Each patent, patent application, publication of a patent application, and other materials, for example, articles, books, instructions, publications, documents, and products, that are cited in this specification and that are other than any document that is inconsistent or conflict with this document or that imposes limiting impact on the widest scope of the claims can be incorporated herein by reference and used for all purposes now or later associated with this document. In addition, the terms in this document shall prevail if there is any inconsistency or conflict between the descriptions, definitions, and/or use of the related terms in any material and the descriptions, definitions, and/or use of the related terms in this document.

[0141] Finally, it should be understood that the principles of the implementation solutions of this specification are described in the implementation solutions of the application disclosed in this specification. Other modified embodiments also fall within the scope of this specification. Therefore, the embodiments disclosed in this specification are merely examples and not limitations. A person skilled in the art can use an alternative configuration according to the embodiments of this specification to implement the application in this specification. Therefore, the embodiments

of this specification are not limited to the embodiments accurately described in the application.

CLAIMS

What is claimed is:

1. A method for training an information extraction model, comprising:

obtaining raw text information corresponding to a sample document and information extraction labeling information corresponding to the sample document, wherein the sample document is a document described in at least one language in a first language set, the raw text information comprises N raw text units in the sample document, and N is an integer greater than or equal to 1;

generating augmented text information corresponding to the sample document based on the raw text information, wherein the augmented text information comprises at least one augmented text unit corresponding to each of at least some of the N raw text units, and the augmented text unit is obtained by translating some text content in the raw text unit into at least one language in a second language set; and

training the information extraction model based on the raw text information, the augmented text information, and the information extraction labeling information.

2. The method according to claim 1, wherein the training the information extraction model based on the raw text information, the augmented text information, and the information extraction labeling information comprises:

pre-training the information extraction model based on the raw text information and the augmented text information, so that a pre-trained information extraction model has a capability of representing a target text unit, wherein the target text unit is a text unit described in any language in the first language set and the second language set; and

performing fine-tuning training on the information extraction model based on the raw text information and the information extraction labeling information, so that an information extraction model obtained after fine-tuning training has a capability of extracting key information.

3. The method according to claim 2, wherein the pre-training the information extraction model based on the raw text information and the augmented text information comprises:

obtaining a plurality of sample pairs based on the raw text information and the augmented text information, wherein one sample pair comprises one raw text unit and one augmented text unit that have a correspondence; and

pre-training the information extraction model through contrastive learning based on the plurality of sample pairs.

4. The method according to claim 3, wherein the pre-training the information extraction model through contrastive learning based on the plurality of sample pairs comprises:

inputting the plurality of sample pairs to the information extraction model, to obtain a plurality of feature vector groups, wherein one feature vector group comprises feature vectors respectively corresponding to two text units in one sample pair; and

updating a parameter of the information extraction model with a training objective of minimizing a difference between two feature vectors in a same feature vector group and maximizing a difference between feature vectors in different feature vector groups.

5. The method according to claim 2, wherein the performing fine-tuning training on the information extraction model based on the raw text information and the information extraction labeling information comprises:

inputting the N raw text units to the pre-trained information extraction model, to obtain feature vectors respectively corresponding to the N raw text units by using the information extraction model, and performing key information extraction based on the feature vectors respectively corresponding to the N raw text units, to obtain an information extraction result; and

updating a parameter of the information extraction model with a training objective of minimizing a difference between the information extraction result and the information extraction labeling information.

6. The method according to claim 1, wherein the augmented text unit is generated in the following manner:

performing word segmentation processing on the raw text unit, to obtain a plurality of text subunits;

determining at least one target text subunit from the plurality of text subunits; and

translating the at least one target text subunit in the raw text unit into at least one language in the second language set, to obtain the augmented text unit.

7. The method according to claim 6, wherein the determining at least one target text subunit from the plurality of text subunits comprises:

determining the at least one target text subunit from the plurality of text subunits based on at least one of occurrence frequencies of the plurality of text subunits in the sample document and

parts of speech of the plurality of text subunits.

8. The method according to claim 7, wherein the determining the at least one target text subunit from the plurality of text subunits based on at least one of occurrence frequencies of the plurality of text subunits in the sample document and parts of speech of the plurality of text subunits comprises:

determining a verb and/or a noun with a quantity of occurrence times greater than or equal to a preset threshold in the plurality of text subunits as the target text subunit.

9. The method according to claim 7, wherein the determining the at least one target text subunit from the plurality of text subunits based on at least one of occurrence frequencies of the plurality of text subunits in the sample document and parts of speech of the plurality of text subunits comprises:

ranking the plurality of text subunits based on a word frequency, and then determining a verb and/or a noun ranked before a preset ranking as the target text subunit.

10. The method according to claim 6, wherein the determining at least one target text subunit from the plurality of text subunits comprises:

randomly selecting at least one text subunit from the plurality of text subunits as the target text subunit.

11. The method according to claim 1, wherein the augmented text information comprises at least one augmented text unit corresponding to each of M of the N raw text units, M is an integer greater than or equal to 1, and a percentage of the M raw text units in the N raw text units is less than or equal to a preset threshold.

12. An information extraction method, comprising:

obtaining a target document;

performing text recognition on the target document, to obtain target text information, wherein the target text information comprises at least one text unit in the target document; and

extracting information from the target text information by using an information extraction model, to obtain key information in the target document, wherein the information extraction model is obtained through training by using the method according to any one of claims 1 to 11.

13. A system for training an information extraction model, comprising:

at least one storage medium that stores at least one instruction set for training an information extraction model; and

at least one processor that is communicatively connected to the at least one storage medium, wherein when the training system runs, the at least one processor reads the at least one instruction set, and performs the following operations based on an indication of the at least one instruction set:

obtaining raw text information corresponding to a sample document and information extraction labeling information corresponding to the sample document, wherein the sample document is a document described in at least one language in a first language set, the raw text information comprises N raw text units in the sample document, and N is an integer greater than or equal to 1;

generating augmented text information corresponding to the sample document based on the raw text information, wherein the augmented text information comprises at least one augmented text unit corresponding to each of at least some of the N raw text units, and the augmented text unit is obtained by translating at least some text content in the raw text unit into at least one language in a second language set; and

training the information extraction model based on the raw text information, the augmented text information, and the information extraction labeling information.

14. The training system according to claim 13, wherein to train the information extraction model based on the raw text information, the augmented text information, and the information extraction labeling information, the at least one processor performs the following operations:

pre-training the information extraction model based on the raw text information and the augmented text information, so that a pre-trained information extraction model has a capability of representing a target text unit, wherein the target text unit is a text unit described in any language in the first language set and the second language set; and

performing fine-tuning training on the information extraction model based on the raw text information and the information extraction labeling information, so that an information extraction model obtained after fine-tuning training has a capability of extracting key information.

15. The training system according to claim 14, wherein to pre-train the information extraction model based on the raw text information and the augmented text information, the at least one processor performs the following operations:

obtaining a plurality of sample pairs based on the raw text information and the augmented text information, wherein one sample pair comprises one raw text unit and one augmented text unit that have a correspondence; and

pre-training the information extraction model through contrastive learning based on the plurality of sample pairs.

16. The training system according to claim 15, wherein to pre-train the information extraction model through contrastive learning based on the plurality of sample pairs, the at least one processor performs the following operations:

inputting the plurality of sample pairs to the information extraction model, to obtain a plurality of feature vector groups, wherein one feature vector group comprises feature vectors respectively corresponding to two text units in one sample pair; and

updating a parameter of the information extraction model with a training objective of minimizing a difference between two feature vectors in a same feature vector group and maximizing a difference between feature vectors in different feature vector groups.

17. The training system according to claim 14, wherein to perform fine-tuning training on the information extraction model based on the raw text information and the information extraction labeling information, the at least one processor performs the following operations:

inputting the N raw text units to the pre-trained information extraction model, to obtain feature vectors respectively corresponding to the N raw text units by using the information extraction model, and performing key information extraction based on the feature vectors respectively corresponding to the N raw text units, to obtain an information extraction result; and

updating a parameter of the information extraction model with a training objective of minimizing a difference between the information extraction result and the information extraction labeling information.

18. The training system according to claim 14, wherein the augmented text unit is generated in the following manner:

performing word segmentation processing on the raw text unit, to obtain a plurality of text subunits;

determining at least one target text subunit from the plurality of text subunits; and

translating the at least one target text subunit in the raw text unit into at least one language in the second language set, to obtain the augmented text unit.

19. An information extraction system, comprising:

at least one storage medium that stores at least one instruction set for information extraction; and

at least one processor that is communicatively connected to the at least one storage medium, wherein when the information extraction system runs, the at least one processor reads the at least one instruction set and implements the information extraction method according to claim 12.

20. A computer-readable nonvolatile storage medium, wherein the computer-readable nonvolatile storage medium stores at least one instruction set, and when the at least one instruction set is executed by at least one processor, the method according to any one of claims 1 to 11 is implemented, or the method according to claim 12 is implemented.