

# **METHOD AND APPARATUS FOR KNOWLEDGE GRAPH CONSTRUCTION, DEVICE AND STORAGE MEDIUM**

## **TECHNICAL FIELD**

[ 0001] This application relates, but is not limited, to the field of natural language processing, and in particular, to a method and apparatus for knowledge graph construction, a device and a storage medium.

## **BACKGROUND**

[ 0002] Events are basic units that describe the activities and states of the objective world. A knowledge graph taking the events as nodes summarizes the events contained in each field in real life and the interactive relationship between the events, which may be very important in application of a series of downstream tasks such as common sense reasoning and the like.

[ 0003] In recent years, there are a lot of related work on constructing event knowledge graphs, such as constructing small-scale and high-quality event knowledge graphs by manual tagging, and constructing a large-scale and relatively-low-quality event knowledge graph by automatic rules. However, it is still an unsolved problem to construct high-quality and relatively-large-scale event knowledge graphs without relying too much on manual tagging. In addition, there is no strict design of a set of ontology for the event knowledge graph and realization of complementation and fusion of multi-modal information (multiple modal information, for example, image and text are respectively one kind of modal information).

## **SUMMARY**

[ 0004] In view of this, embodiments of this application at least provide a method and

apparatus for knowledge graph construction, a device and a storage medium.

[ 0005]       The technical solutions of the embodiments of this application are implemented as follows.

[ 0006]       In a first aspect, the embodiments of this application provide a method for knowledge graph construction, which may include that: event detection is performed on text data to obtain at least one trigger word of at least one first instance event corresponding to the text data; image classification processing is performed on image data to obtain at least one event type of at least one second instance event corresponding to the image data; a first concept event matched with each trigger word is determined based on a multi-modal ontology; a second concept event matched with each event type is determined based on the multi-modal ontology; and a multi-modal event knowledge graph is constructed based on the first concept event matched with each trigger word and the second concept event matched with each event type.

[ 0007]       In a second aspect, the embodiments of this application provide a method for multi-modal ontology construction, which may include that: at least one word sense in a lexicon is matched with each event in a first frame respectively, a first mapping relation between the lexicon and the first frame is established, and the first frame is configured to determine a concept event of text data; the at least one word sense in the lexicon is matched with each event in a second frame respectively, a second mapping relation between the lexicon and the second frame is established, and the second mapping relation is configured to determine a concept event of image data; and the lexicon, the first frame and the second frame are fused based on the first mapping relation and the second mapping relation to obtain the multi-modal ontology.

[ 0008]       In a third aspect, the embodiments of this application provide a knowledge graph construction apparatus, which may include: a detection module, configured to perform event detection on text data to obtain at least one trigger word of at least one first instance event corresponding to the text data; a classification module, configured to perform image classification processing on image data to obtain at least one event type of at least one second instance event corresponding to the image data; a first determination module, configured to determine a first concept event matched with each trigger word

based on a multi-modal ontology; a second determination module, configured to determine a second concept event matched with each event type based on the multi-modal ontology; and a construction module, configured to construct a multi-modal event knowledge graph based on the first concept event matched with each trigger word and the second concept event matched with each event type.

[ 0009]        In a fourth aspect, the embodiments of this application provide a multi-modal ontology construction apparatus, which may include: a first establishment module, configured to match at least one word sense in a lexicon with each event in a first frame respectively, and establish a first mapping relation between the lexicon and the first frame, the first frame is configured to determine a concept event of text data; a second establishment module, configured to match the at least one word sense in the lexicon with each event in a second frame respectively, and establish a second mapping relation between the lexicon and the second frame, the second frame is configured to determine a concept event of image data; and a fusion module, configured to fuse the lexicon, the first frame and the second frame based on the first mapping relation and the second mapping relation to obtain the multi-modal ontology.

[ 0010]        In a fifth aspect, the embodiments of this application provide a computer device, which may include a memory and a processor. The memory stores a computer program executable on the processor. The processor, when executing the computer program, implements part of all of steps in the above method.

[ 0011]        In a sixth aspect, the embodiments of this application provide a computer-readable storage medium. The computer-readable storage medium stores a computer program. Part or all of steps in the above method are implemented when the computer program is executed by a processor.

[ 0012]        In a seventh aspect, the embodiments of this application provide a computer program. The computer program includes computer-readable codes. When the computer-readable codes are operated in a computer device, a processor in the computer device executes and implements part or all of steps in the above method.

[ 0013]        In an eighth aspect, the embodiments of this application provide a computer

program product. The computer program product includes a non-transitory computer-readable storage medium storing a computer program. When the computer program is read and executed by a computer, part or all of steps in the above method are implemented.

[ 0014]        In the embodiments of this application, the first concept event of the text data and the second concept event of the image data are determined based on the multi-modal ontology, and then the multi-modal event knowledge graph is constructed based on the first concept event and the second concept event, so that complementation and fusion of the multi-modal information are realized, more information is provided for construction of the knowledge graph, and more possibilities are provided for knowledge mining.

[ 0015]        It should be understood that, the above general description and the following detailed description are merely exemplary and explanatory, and do not limit this application.

#### BRIEF DESCRIPTION OF THE DRAWINGS

[ 0016]        Drawings are incorporated into the specification and constitute a part of the specification. The drawings illustrate embodiments in accordance with this application and serve to describe the technical solutions of this application together with the specification.

[ 0017]        FIG. 1 is a schematic diagram of an example of text and image events and a relation between the text and image events according to an embodiment of this application.

[ 0018]        FIG. 2 is a schematic diagram of three different grades in a Multi-modal Event Knowledge Graph (MMEKG) from left to right according to an embodiment of this application.

[ 0019]        FIG. 3 is a schematic diagram of an ontology designed in a MMEKG according to an embodiment of this application.

- [ 0020] FIG. 4A is a schematic diagram of a knowledge extraction system according to an embodiment of this application.
- [ 0021] FIG. 4B is a schematic diagram of a pairing example constructed for concept event induction and relation induction according to an embodiment of this application.
- [ 0022] FIG. 5A is a schematic implementation flowchart of a method for knowledge graph construction according to an embodiment of this application.
- [ 0023] FIG. 5B is a schematic implementation flowchart of S104 in a method for knowledge graph construction according to an embodiment of this application.
- [ 0024] FIG. 5C is a schematic implementation flowchart of S105 in a method for knowledge graph construction according to an embodiment of this application.
- [ 0025] FIG. 5D is a schematic implementation flowchart of adding an event relation into a MMEKG according to an embodiment of this application.
- [ 0026] FIG. 5E is a schematic implementation flowchart of S107a in FIG. 5D.
- [ 0027] FIG. 5F is a schematic implementation flowchart of an ontology updating method according to an embodiment of this application.
- [ 0028] FIG. 5G is a schematic implementation flowchart of a method for multi-modal ontology construction according to an embodiment of this application.
- [ 0029] FIG. 5H is a schematic implementation flowchart of a multi-modal ontology updating method according to an embodiment of this application.
- [ 0030] FIG. 6A is a schematic diagram of a composition structure of an apparatus for knowledge graph construction according to an embodiment of this application.
- [ 0031] FIG. 6B is a schematic diagram of a composition structure of an apparatus for multi-modal ontology construction according to an embodiment of this application.
- [ 0032] FIG. 7 is a schematic diagram of a hardware entity of a computer device according to an embodiment of this application.

## DETAILED DESCRIPTION

- [ 0033] For making the purposes, technical solutions and advantages of this application clearer, the technical solutions of this application will further be described below in combination with the drawings and embodiments in detail. The described embodiments should not be considered as limits to this application. All other embodiments obtained by those of ordinary skill in the art without creative work shall fall within the scope of protection of this application.
- [ 0034] "Some embodiments" involved in the following descriptions describes a subset of all possible embodiments. However, it can be understood that "some embodiments" may be the same subset or different subsets of all the possible embodiments, and may be combined without conflicts.
- [ 0035] Term "first/second/third" involved is only for distinguishing similar objects and does not represent a specific sequence of the objects. It can be understood that "first/second/third" may be interchanged to specific sequences or orders if allowed to implement the embodiments of this application described herein in sequences except the illustrated or described ones.
- [ 0036] Further, please note that the term "A matched with each B" involved in the following descriptions indicates a respective one of at least one A matched with each of at least one B. For example, a first concept event matched with each trigger word indicates a respective one of at least one first concept event matched with each of at least one trigger word.
- [ 0037] Unless otherwise defined, all technical and scientific terms in the specification have the same meaning as those skilled in the art, belonging to this application, usually understand. Terms used in the specification are only used for describing the purpose of this application, but not intended to limit this application.
- [ 0038] The embodiments of this application provide a large-scale Multi-modal Event Knowledge Graph (MMEKG). The MMEKG unifies different knowledge models

through events, and complements each other and disambiguates each other. Herein, the MMEKG includes: (i) more than 0.99 million concept events and 644 relation types (hierarchical and semantic relations between the concept events) to cover most types of events. (ii) More than 863 million instance events connected by 934 million relations (relations between the instance events, which may be further divided into more granular subtypes such as temporal, causal, co-occurrence, and other semantic relations) that provide rich contextual information in text and/or images.

[ 0039]       The embodiments of this application also provide an efficient text/visual knowledge extraction system pipeline, configured to collect billions of instance events and relations between them. The embodiments of this application also provide an induction strategy to create million-scale concept events and ontologies configured to organize all events and the relations between the events in the MMEKG. The embodiments of this application also provide a pipeline which may seamlessly parse the text/image into an event graph and retrieve multi-modal knowledge at the concept and instance levels.

[ 0040]       Recently, many Knowledge Graphs (KGs) have been collated and successfully applied in a variety of scenarios, from information extraction to information retrieval. The KG generally stores billions of world facts in a directed graph. Herein, nodes represent entities and edges represent the relations between them. Although simple but efficient, this entity-centric KG has limited expression capacity. How to represent more complex knowledge, such as events, situations or different modalities, becomes a key question for wider application of the KG.

[ 0041]       Based on this, the embodiments of this application provide the large-scale MMEKG that may connect, supplement and eliminate different forms of knowledge for better understanding or reasoning. Similar to what happens in the real world, the MMEKG takes the events as its basic components, each defined by a concept, several arguments, and a corresponding role. There are various types of relations between the events, such as causal relations, temporal relations or sub-event relations, and the entities are arguments in the KG.

[ 0042]       FIG. 1 shows an example of text and image events and a relation between the

text and image events. The text and image events are a visual sleep event with arguments cat (sleeper) and soft (place), and a textual dressing event with arguments cat (wear) and scarf (clothing) respectively. Herein, CO-REF represents Co-reference, that is, a relation of co-reference, and an argument role is located in a parenthesis. The two events not only connect the text and image with complementary arguments, but also provide a potential common sense of covering with a scarf while sleeping.

[ 0043]        Compared with the event KG in the related art, the MMEKG promotes the development of the field in the following three aspects: (1) a large ontology includes 0.99 million concept events and 644 relation types, which cover most types of events in the real world. (2) The multi-modal knowledge is naturally fused. The KG realizes fusion of different modal data through granular alignment of events and arguments for the first time. (3) The concept events and the instance events are integrated, so that the ontology may be expanded from the instance events, and the common sense knowledge at the concept level may be obtained through comprehensive reasoning provided by the context.

[ 0044]        The operation of constructing the MMEKG mainly includes the following two steps that: (1) in order to construct a mode and acquire the concept event, firstly, FrameNet and WordNet are manually combined to initialize a high-quality event ontology; and then the ontology is automatically expanded by inductive reasoning of the instance event. In the embodiments of this application, a Simple Event Model (SEM) is expanded to define the ontology provided in the embodiments of this application in a Resource Description Framework (RDF). (2) In order to extract the instance event from the text or image, the embodiments of this application develops a knowledge extraction system to support rapid and large-scale extraction in an actual scene. The system includes event extraction and event relation extraction under two modalities, and alignment between the two modalities. In addition, the system may parse the input text/image into the event graph and seamlessly retrieve the multi-modal knowledge from the MMEKG.

[ 0045]        The KG provided in the embodiments of this application is described in detail below.

[ 0046]        Part 1: Introduction to MMEKG. As FIG. 2 shows three different grades in the MMEKG from left to right, the left is an extracted multi-modal content, the middle

shows an instance event integrated from the original content, and the right is reasoned concept events. The MMEKG provided in the embodiments of this application is different from a conventional event-centric KG. Herein, the MMEKG includes four types of nodes and four types of relations. The nodes include: a concept event, an instance event, an entity argument and a non-entity argument (such as a character). The four types of relations include: (1) a relation between the instance events, which may be further divided into more granular subtypes such as temporal, causal, co-occurrence and other semantic relations. (2) A relation between the concept events, called subclassOf, represents a hierarchical relation. (3) A relation between the concept event and the instance event, called instanceOf and configured to integrate the concept event and the instance event. (4) A role relation which reflects a role of an associated event argument (entity or non-entity). Herein, different concept events have different roles.

[ 0047] Definition is as follows:  $MMEKG = \{(h, r, t) | h, t \in \mathcal{E}, r \in \mathcal{R}\}$ ;  $\mathcal{E} = \mathcal{E}_{cpt} \cup \mathcal{E}_{ins} \cup \mathcal{E}_{ent} \cup \mathcal{E}_{nent}$ ,  $\mathcal{R} = \mathcal{R}_{ins-ins} \cup \mathcal{R}_{cpt-cpt} \cup \mathcal{R}_{cpt-ins} \cup \mathcal{R}_{role}$ .

[ 0048]  $\mathcal{E}_{cpt}$  represents a concept event,  $\mathcal{E}_{ins}$  represents an instance event,  $\mathcal{E}_{ent}$  represents an entity, and  $\mathcal{E}_{nent}$  represents a non-entity.  $\mathcal{R}_{ins-ins}$  represents a relation between the instance events,  $\mathcal{R}_{cpt-cpt}$  represents a relation between the concept events,  $\mathcal{R}_{cpt-ins}$  represents a relation between the concept event and the instance event, and  $\mathcal{R}_{role}$  represents a role relation.  $W(h, r, t)$  represents the weight of a triple  $(h, r, t)$  relation, expressed with confidence, and  $h$ ,  $r$ , and  $t$  are a head, a relation and a tail in the triple respectively.

[ 0049] Part 2: User interface and system architecture. Based on the MMEKG and the text/visual knowledge extraction system, the embodiments of this application develop a prototype system that may parse the text or image into the event graph. Herein, the nodes represent entities and the edges represent the relations between them. Each instance event may refer to specific information: arguments corresponding to the concept event, a synonym and the role, and connected neighbors in the MMEKG.

[ 0050] For each instance event, the embodiments of this application identify a trigger word and a synonym thereof, and link the trigger word and the synonym to the concept

event in the MMEKG. Then, a corresponding role is determined according to the concept event and the argument is extracted in the text/image.

[ 0051]       The text/visual knowledge extraction system includes three main components: text knowledge extraction, visual knowledge extraction and event relationship extraction. The event relation extraction is configured to extract various relations between the events and includes fusion of text and visual events. It is to be noted that the concept event, the synonym and the relation type are defined by the cross-modal event ontology provided in the embodiments of this application.

[ 0052]       Part 3: Introduction to the multi-modal ontology. The ontology is crucial, not only limiting the type of knowledge of interest, but also providing reasoning capabilities, that is, the induction from the instance event to the concept event brings new knowledge. The establishment of the ontology includes that: (1) ontology data is modelled; (2) an initial ontology is established by combining external resources; and (3) ontology induction is performed by continuous expansion.

[ 0053]       The above three steps are described below respectively.

[ 0054]       (1) The ontology data is modelled.

[ 0055]       In the embodiments of this application, the basic SEM is inherited and expanded as a basis for knowledge representation. FIG. 3 is a schematic diagram of an ontology designed in an MMEKG. Herein, a dashed box represents the character, a solid box represents the event, entity and relation, a box filled in the latticed shape represents event correlation, a box filled in the gray shape represents relation correlation, a box filled in the dotted shape represents text/image information extracted from the instance event, and a non-filled circle represents a virtual node connecting the instance event and its source information.

[ 0056]       1. The single event is expanded from the SEM and FrameNet, that is, the RDF structure of the single event is expanded from the definition of the SEM and the FrameNet. As shown in FIG. 3, each role has an associated ekg:[role] connecting the instance event and the argument, rdf: refers to a type defined by the rdf ontology, sem: refers to a type defined by the SEM ontology, and ekg: refers to a new type defined by

the embodiments of this application.

[ 0057]        2. The virtual node is added. Herein, the virtual node is connected with the instance event through `ekg:contextOf` to represent the source of this instance event. The edges of the virtual node, such as `ekg:trigger`, `ekg:modality` and `ekg:content`, represents the trigger word, modality, and sentence/image index of the instance event respectively. The virtual node is a node designed for clear structure and needed for local structure integration, and the virtual node is queried only when needing to be queried, so that unnecessary query is reduced.

[ 0058]        3. The relation between the events is constructed, which mainly includes the following three relations.

[ 0059]        The first relation is `rdf:instanceOf`, that is, the relation between the instance event and the concept event.

[ 0060]        The second relation is `rdf:subclassOf`, that is, the relation between the concept events, which is configured to represent the hierarchical relation between the concept events.

[ 0061]        The third relation is other various relations between the instance events, such as timing or causal relation, etc. Based on the relation, a link node is added between the instance events to represent the relation between the instance events. As shown in FIG. 3, a label `[EventRel_id]` is a link node, configured to indicate the relation between `ins-Event_317` and other events (not shown in the figure), and middle brackets indicate that the content in the brackets may be replaced. Therefore, in a case that a plurality of event relations exist between two instance events, different event relations may be represented through different link nodes, so that different event relations may be quickly queried (only the link nodes corresponding to the required event relations need to be queried) and relations between the instance events may be quickly modified (expression of the edges between the instance events does not need to be modified, and only information of the link nodes may be modified), and the ontology has good expandability. For example, the event relation has a plurality of major types such as temporal relations, the major type includes minor types such as before and after, one `[Reltype id]` represents one relation,

one minor type corresponds to one [Reltype id], and in a case that the event information with the time relation being before needs to be queried, the event node with the time relation being after maybe directly deleted through the [Reltype id], so that quick query is realized.

[ 0062]        4. Information statistics and supplementation are integrated. For example, the frequency and a significance score are added to reasoning of uncertainty events. For example, the instance events are: "fry\_egg\_in the morning" and "fry\_egg\_with a scoop". Herein, the frequency of occurrence of "fry\_egg" is high and the significance score is high, so that the finally reasoned concept event is "fry\_egg", which may be described in the ontology induction part in detail below.

[ 0063]        (2) The initial ontology is established by combining the external resources.

[ 0064]        In the embodiments of this application, the ontology is initialized by combining WordNet, FrameNet, and imSitu ontologies.

[ 0065]        1. A synonym set of each verb and adjective in the WordNet is mapped to one frame in the FrameNet (such as roast.v.01→Apply\_heat). Herein, the frame is a high-level concept event, one frame represents one concept event, due to wide coverage of the event in the FrameNet, for example, one chemical substance is converted to another called Cause\_change, and changing name is also called Cause\_change, while the WordNet has richer word sense, the aligned synonym set may make the concept event more granular. In addition, the WordNet also brings hierarchical information to the FrameNet. By aligning the WordNet with the FrameNet, it is possible to know what event is described once the word sense of the trigger word of the text or image is known, for example, in the case where Fire indicates firing, it corresponds to an attack event in the FrameNet and in the case where Fire indicates termination of employment it corresponds to an event of loss of work in the FrameNet.

[ 0066]        For mapping, the embodiments of the application first consider a structure mapping result jointly, that is, in previous work, a mapping relation between part of word senses in the WordNet and the frame in the FrameNet is given, and the embodiments of this application directly adopt the mapping result. For the word sense without the

mapping relation in the WordNet, the mapping relation is determined by solving a cosine similarity score between the word sense in the WordNet and the event type of the frame, and the matching relation with the highest cosine similarity is selected for mapping.

[ 0067] In the embodiments of this application, 100 pairs of synset and frame pairs are randomly extracted to check whether the definitions of the mapped synset and frame are aligned or not, and finally 89% are found to be reasonable.

[ 0068] 2. During implementation, the word sense in the WordNet may be manually aligned with the event in the Imsitu. Due to the fact that the event type of the event in the Imsitu is marked, the WordNet may be aligned with the Imsitu based on the word sense in the WordNet and the event type of the event in the Imsitu, and the association between the WordNet and the Imsitu is realized. Due to the fact that the Imsitu is configured to describe the image event, an image event triggering event is triggered by a verb, and the verb corresponding to the event type in the Imsitu has no word sense, the verb needs to be aligned with the word sense under the WordNet, and due to the fact that the Wordnet is aligned with the Framenet, the event type under the Framenet may be known through the Wordnet, so that the concept event corresponding to the image is obtained.

[ 0069] (3) Ontology induction is performed by continuous expansion.

[ 0070] 1. Induction is performed on the concept event.

[ 0071] At S1, assuming an initialized concept event  $o$  and a specific role  $r$  of the concept event  $o$ , all arguments under the specific role  $r$  are selected and are connected with an instance event  $O_{in}$  through  $r$ . Herein, the instance event  $O_{in}$  is classified as the concept event  $o$ .

[ 0072] At S2, word form reduction and heuristic clustering are performed on head words of all the arguments. The word form reduction is to remove affixes of the words and extract the trunk parts of the words, and the extracted words are generally words in a dictionary. For example, the initialized concept event is fry, the specific role is food, all the instance events for fry are found, and the arguments in the food role are counted: fish, a fish, the fish, a big fish, chicken and the chicken. Herein, fish, a fish, the fish and a big fish are all fish after the word form reduction, chicken and the chicken are all chicken

after the word form reduction, and two types of fish and chicken are obtained after heuristic clustering is performed.

[ 0073]        At S3, a significance score of each type is calculated, and the following two aspects are considered: (1) a confidence score (the times of occurrence) of each type, and (2) importance of each type, for example, the type is "I", then the importance is not high, and the weight of the type may be reduced.

[ 0074]        At S4, K types with the highest significance scores are selected and a new concept event is created by combining the role r and these names with the trigger word. For example, the times of occurrence of fish is higher, the significance score is high, when K is 1, fish is combined with the trigger word fry and the role food, and a new concept event "fry fish" is derived.

[ 0075]        As shown in FIG. 2, according to the embodiments of this application, the new concept event is derived, such as "complete.v.01\_Activity:tour" and "complete.v.01\_Activity:Time:fall.". The granular concept events are obtained by summarizing the instance events through the instanceOf relation and summarizing complete.v.01 and the subclass relation.

[ 0076]        2. Induction is performed on the concept event relation.

[ 0077]        The purpose of relation induction is to discover a common sense relation between the concept events based on the relation between the instance events. During implementation, the significance score  $sr(o_h, o_t)$  of each pair of instance events  $(o_h, o_t)$  on r is calculated to determine whether it is induced to be the relation between the concept events. Herein, the calculation of the significance score considers the following two aspects: (1) the confidence score of the relation r between the instance events, and (2) the times of occurrence of a tail event (as little as possible). In a case that the significance score exceeds a threshold of the MMEKG, the relation of the instance event  $(o_h, o_t)$  is added to the relation of the concept event. For example, in FIG. 2, due to the fact that the significance score of the triple (talk.v.01, co-occur and sit.v.01) exceeds the threshold, the relation is expanded from the instance level to the concept level in the embodiments of this application. Subscripts of  $o_h$  and  $o_t$  represent a head event and the tail event in the

triple corresponding to each pair of instance events respectively.

[ 0078] In the embodiments of this application, through the concept event induction and the concept event relation induction, a more hierarchical and granular classification structure is realized.

[ 0079] Part 4: Knowledge extraction system.

[ 0080] A knowledge extraction system collects large-scale instance events and relations for the MMEKG. The embodiments of this application follow the overall framework of previous knowledge extraction systems, but expand and optimize event-related components to enable them to efficiently extract billion-scale high-quality events. With more advanced models, adjustment strategies and component architectures, the system achieves the same, if not better, performance on each component using a common benchmark test. In some embodiments, all cross encoders in the system may be replaced with dual encoders and a multi-tasking trained joint model is executed during event relation extraction to improve efficiency.

[ 0081] FIG. 4A shows a schematic diagram of a knowledge extraction system according to an embodiment of this application. As shown in FIG. 4A, the knowledge extraction system mainly includes four components: a cross-modal event ontology 43, a text knowledge extraction component 41, a visual knowledge extraction component 42 and an event relation extraction component 44. The introduction to the cross-modal event ontology is described above, and the other three parts are described below.

[ 0082] 1. Text Knowledge Extraction Component

[ 0083] The operation that the text knowledge extraction component extracts the node of the event graph from the unstructured text by event detection and argument extraction includes the following steps.

[ 0084] At S1, a corpus is preprocessed. Boundaries of a document (5 to 10 sentences per document) are identified using a Next Sentence Prediction (NSP) model based on a pre-trained language representation model BERT (Bidirectional Encoder Representation from Transformers) and heuristic rules. Then, a Part Of Speech-tag (POS-tag) and a

dependency tree are obtained by an entity naming recognition tool (such as Stanza). Herein, the dependency tree is derived by Dependency Parsing, and a candidate trigger word of the text is determined. The verb and the adjective are considered to be candidate trigger words.

[ 0085]       At S2, event detection is converted into an unsupervised Word Sense Disambiguation (WSD) task based on the synonym set in the ontology to avoid expensive training data. A Biencoder model is applied to predict the most probable synonym set of the candidate trigger word. Due to the fact that each synset refers to one concept event, the text is connected with the MMEKG.

[ 0086]       At S3, the embodiments of this application provide an effective Event Argument Extraction (EAE) method, marked as Prompting Argument Interaction for Event Argument Extraction (PAIE). Its basic idea is to expand a Question-Answer (QA)-based model to predict all roles of a target event at the same time. The embodiments of this application provide a Pre-trained Language Model (PLM) which is configured to extract tasks and design a role interaction prompt template for each concept event. All roles embedded with text vectors act as query vectors to determine argument spans as answers, so that the arguments are obtained. The model is trained on annotations provided by the FrameNet.

[ 0087]       2. Visual Knowledge Extraction component

[ 0088]       For the visual knowledge extraction component, a two-stage extraction network is adopted. The two models are trained using the largest visual situation recognition data set.

[ 0089]       (1) A pre-trained vision model Vision Transformer (ViT) is utilized to perform motion recognition to obtain image block features, image classification is performed, and the concept event is determined based on the image classification result.

[ 0090]       (2) A pre-trained residual network such as ResNet 50 is used as a backbone network of Faster Region-CNN (Faster R-CNN), and a conditional Long Short-Term Memory (LSTM) decoder is configured to aggregate role information and extract the argument from the image.

[ 0091]       The ViT model is modified from the conventional Transformer structure, which includes an Encoder-Decoder framework, but for the ViT, only the encoder part is used. The input of the standard Transformer is one-dimensional sequence data, so that the image needs to be converted into sequence data. The idea of the ViT is to divide the image into fixed-size patches without overlapping, and then convert each patch into a one-dimensional vector through stretching operation. Finally, the input patches are transformed into a fixed-length vector through a linear transformation layer, which is called patch\_embedding.

[ 0092]       3. Event Relation Extraction component

[ 0093]       The event relation extraction component is designed to extract temporal, causal, co-occurrence and semantic relations between the instance events. Herein, co-occurrence includes text/image alignment.

[ 0094]       (1) Temporal and causal relations. The embodiments of the application provide a new method, that is, for constructing a document-level graph to infer the relation between the events globally. In the absence of clear temporal/causal indicators and complex heuristic rules, the method in the embodiments of this application may perform cross-sentence reasoning, which enables the embodiments of this application to simultaneously and effectively identify all temporal and causal relations of the document. In the embodiments of this application, the temporal and causal relations are jointly predicted into multi-label and multi-task classification, and a model is trained based on a causal time library and includes six relation types: before, after, during, including, included and causal relation.

[ 0095]       (2) Text co-occurrence. If a conjunctive relation exists between the trigger words, text co-occurrence is considered. In the embodiments of this application, the conjunctive relation is identified through dependency parsing.

[ 0096]       Cross-modal co-occurrence. In the embodiments of this application, the events are extracted from the paired image characters respectively, and it is assumed that the events co-occur, so that mining of information between the paired image characters is achieved. The embodiments of this application also observe semantic conversion

between different modalities. As shown in FIG. 1, the text dressing event may be a sub-event of a visual sleep event.

[ 0097] (3) Semantic relations

[ 0098] In the embodiments of this application, when the argument of an event A is a gerund phrase B, B may also be considered a sub-event of A. Herein, the event A is triggered by the gerund as its semantic element. For example, two events are extracted from this sentence "eating too much fried chicken causes overweight": causing overweight (event A) and eating too much chicken (event B). Due to the fact that A is also an argument of the role influencing\_entity of B, the events of eating too much chicken and causing overweight are connected by the relation influencing\_entity. Based on such assumption, the relation type is expanded using the argument in the FrameNet.

[ 0099] An event pair satisfying the following two conditions is defined to be in a semantic relation: (1) in dependency parsing, the trigger words are connected through csubj (subject clause), xcomp (open complement clause), ccomp (complement clause) or advcl (adverbial clause); (2) the trigger word of one event is the argument of another event.

[ 00100] Part 5: Evaluation

[ 00101] (1) MMEKG

[ 00102] Table 1 shows statistical data of the MMEKG and other event KGs. In the embodiments of this application, a full version (MMEKG full) and a core version (MMEKG Core) are built, as shown in Table 1, the MMEKG Core filters out infrequently occurring events (less than 3 times) to obtain a denser and more accurate version. The MMEKG involves not only larger ontologies, but also more instance events.

Table 1

|            | Instance Event | Concept Event | Relation   | Relation Type |
|------------|----------------|---------------|------------|---------------|
| ConceptNet | —              | 74,989        | 116,097    | 4             |
| ATOMIC     | —              | 309,515       | 877,108    | 8             |
| ASER core  | 52,940,258     | —             | 52,296,498 | 14            |

|            |             |         |             |     |
|------------|-------------|---------|-------------|-----|
| ASER full  | 438,648,952 | —       | 648,514,465 | 14  |
| MMEKG core | 12,310,716  | 990,123 | 48,599,695  | 644 |
| MMEKG full | 863,428,946 | 990,123 | 934,413,371 | 644 |

[ 00103]        Herein, ConceptNet, ATOMIC and ASER are event common sense knowledge bases in the field of KGs as a contrast, ASER full is a full version of ASER, and ASER core is a core version of ASER.

[ 00104]        (2) Text/visual extraction system performance

[ 00105]        Table 2 shows results of components trained on a public dataset in the embodiments of this application due to the fact that there is no uniform benchmark to evaluate the entire extraction process. It is to be seen that all of the knowledge extraction components in the embodiments of this application achieve better performance except for the WSD. The WSD model provided in the embodiments of this application has high performance in the aspect of large-scale event detection.

Table 2

| Component                   | Sub-task                             | Benchmark       | Metric   | Score |
|-----------------------------|--------------------------------------|-----------------|----------|-------|
| Text knowledge extraction   | WSD                                  | AemEval-2007    | Accuracy | 74.5  |
|                             | EAE                                  | ACE-05          | F1       | 67.0  |
| Visual knowledge extraction | Verb Detection (VerbD)               | imSitu          | Accuracy | 46.8  |
|                             | EAE                                  | imSitu          | Accuracy | 23.8  |
| Event relation extraction   | Event Causality Identification (ECI) | Causal-TimeBank | F1       | 61.7  |

[ 00106]        (3) Instance-level evaluation

[ 00107]        Considering the difference in data distribution between the training data and the extracted corpus, the instance-level quality of the MMEKG may be manually evaluated during implementation. 1000 instance events were randomly selected from the text and 500 instance events were randomly selected from the image. In addition to the

original context, the embodiments of this application invite many people to mark whether the extracted event represents the semantics of the original source. For the instance event relations, the embodiments of this application consider: (1) causal/temporal relations from the text and (2) cross-modal co-occurrence from image title pairs. 200 text relations and 300 cross-modal relations were sampled. In addition to the context, these extracted relations were provided for the same people and asked if the extracted relations are matched with the original resources. The results in Table 3 show that there is little drop in performance in consideration of the complexity of the overall pipeline, and the quality of the MMEKG provided in the embodiments of this application may be acceptable.

Table 3 Instance-level evaluation

|        | Modality    | Precision |
|--------|-------------|-----------|
| Event  | Texual      | 84.0%     |
| Event  | Visual      | 64.6%     |
| Triple | Texual      | 66.9%     |
| Triple | Cross-modal | 63.8%     |

[ 00108] (4) Ontology-level evaluation

[ 00109] A large-scale ontology is a key to knowledge inductive reasoning. In the embodiments of this application, the quality of concept event induction and relation induction is further evaluated. Unlike the instance-level evaluation, no context is provided for reference in the ontology evaluation. In order to facilitate comparison, a pairing of one positive sample and one negative sample is constructed in the embodiments of this application. FIG. 4B shows a pairing example constructed for concept event induction (top) and relation induction (bottom). Herein, each pair includes one positive sample and one negative sample. The positive sample is an example extracted from the induced concept event or relation. The negative sample is generated by replacing the argument (concept event induction) or tail event (relation induction) in the positive sample. The same person is queried which sample conforms to the common sense better. The results are shown in Tables 4 and 5, and two negative effects are about 20%. In particular, for relation induction, it is difficult to distinguish which one in some similar pairs is better. The embodiments of this application attribute this to low recall rates and random negative sampling, which may bring false negativity and also provide

insights into future improvements.

Table 4 Concept event induction

| Sample | Positive | Negative |
|--------|----------|----------|
| 1000   | 80.1%    | 19.9%    |

Table 5 Relation induction

| Type     | Sample | Positive | Similar | Negative |
|----------|--------|----------|---------|----------|
| Temporal | 134    | 65.7%    | 15.7%   | 18.6%    |
| Co-occur | 139    | 57.6%    | 20.1%   | 22.3%    |
| Semantic | 137    | 46.0%    | 36.5%   | 17.5%    |
| All      | 550    | 58.5%    | 22.4%   | 19.1%    |

[ 00110] Existing event KGs often face a dilemma of quality and quantity. The embodiments of this application not only develop a larger KG through a larger corpus and an advanced extraction system, but derive a complex ontology, and integrate cross-modal information to control the quality of the KG.

[ 00111] A previous multi-modal knowledge extraction system of the knowledge extraction system jointly extracts small-domain information from relatively small resources. The knowledge extraction system provided by embodiments of this application inherits their overall framework, but is configured to extract billion-scale and global events. Therefore, the event-related components are optimized in a targeted manner in the embodiments of this application, so that the efficiency and the effectiveness are improved.

[ 00112] Some previous work of aligning cross-media event parameters also connects the text and the image through granular alignment of event parameters for various tasks, such as multi-modal event extraction and false news detection. Instead, embodiments of this application fuse knowledge from different modes to build such a large-scale KG.

[ 00113] The embodiments of this application provide a first MMEKG with a large-scale event ontology, which not only communicates and supplements different knowledge modes through more expressive events, but also facilitates comprehensive

reasoning with rich cross-modal contexts. In addition, the embodiments of this application also provide a demonstration system, which may seamlessly parse and link any text/image through the knowledge extraction system.

[ 00114]       The embodiments of this application provide a KG construction method. The method may be executed by a processor of a computer device. The computer device may refer to a server, a laptop computer, a tablet computer, a desktop computer, a smart television, a set-top box, a mobile device (such as a mobile phone, a portable video player, a personal digital assistant, a dedicated messaging device, a portable gaming device), or other devices with data processing capabilities. The method includes the following operations.

[ 00115]       FIG. 5A is a schematic implementation flowchart of a method for knowledge graph construction according to an embodiment of this application. As shown in FIG. 5A, the method includes S101 to S105 as follows.

[ 00116]       At S101, event detection is performed on text data to obtain at least one trigger word of at least one first instance event corresponding to the text data.

[ 00117]       Herein, the text data refers to data in the form of text, and the instance event refers to an event that specifically describes a practical instance. The constituent elements of the event may include a trigger word, an event type, an argument and a role. Herein, the trigger word refers to the core word of the event, and is mostly a verb and an adjective, such as attack, discussion and the like. The event type refers to a category of the event, such as a motion event, an attack event and the like. The argument refers to a participant of the event, such as Zhang San (that is, a specific person), a school and the like. The role refers to a role of the argument in the event, such as an attacker, a location, time and the like.

[ 00118]       The following events are described as an example: The private Dogan news agency reported that more than 30 of those wounded in A, primarily children, were transferred to a hospital in B.

[ 00119]       It showed footage of ambulances arriving at C hospital and medical personnel <t> unloading </t> a girl on stretchers.

[ 00120]       The meaning is that the private Dogan news agency reported that more than 30 of those wounded in A, primarily children, were transferred to a hospital in B. The report showed footage of ambulances arriving at C hospital and medical personnel unloading a girl on stretchers.

[ 00121]       Herein, the event type is a motion event triggered by the trigger word "unloading". The arguments are "A", "ambulances", "C hospital", "medical personnel" and "a girl". Each argument plays a specific role in this event, i.e., an origin (that is, a location indicated by A), a vehicle (that is, a vehicle indicated by ambulances), a destination (that is, a location indicated by C hospital), a transporter (that is, a transporter indicated by the medical personnel), and a passenger (that is, a passenger indicated by a girl), respectively

[ 00122]       In some embodiments, the implementation of S101 may perform POS tagging and Dependency parsing on the text data, and then obtain the trigger word of the text data corresponding to at least one first instance event based on a specific rule. In some embodiments, the specific rule refers to a vast majority of verbs other than stop words and adjectives that satisfy a plurality of dependencies.

[ 00123]       In some embodiments, in a case that the text data is the document, the implementation of S101 may preprocess the text data first, for example, the boundaries of the document (5 to 10 sentences per document) are identified using the NSP model based on the BERT and the heuristic rules. Then, POS tagging and dependency parsing are performed on the preprocessed text data to obtain the at least one trigger word of at least one first instance event corresponding to the text data.

[ 00124]       The embodiments of this application do not limit the method for obtaining the at least one trigger word of at least one first instance event corresponding to the text data

[ 00125]       At S102, image classification processing is performed on image data to obtain at least one event type of at least one second instance event corresponding to the image data.

[ 00126]       Herein, the image data refers to data in the form of an image. The implementation of S102 may adopt a deep learning-based image classification network

such as ViT, Resnet, etc., to perform image classification processing on the image data to obtain the at least one event type corresponding to at least one second instance event of the image data. Herein, in order to enable the image classification network to identify the event type of the second instance event, the image classification network may be trained by adopting the data sample set marked with the event type in advance, so that the image classification network has the capability of identifying the event type. The embodiments of this application do not limit the type of the image classification network.

[ 00127]       At S103, a first concept event matched with each trigger word is determined based on a multi-modal ontology.

[ 00128]       Here, multi-modality means having at least two modalities, such as a text modality and an image modality, or other modalities. The embodiments of this application do not limit the forms of the modalities. The concept event refers to an abstract event summarized by the instance event.

[ 00129]       In some embodiments, the multi-modal ontology may include a lexicon, a first frame and a second frame. Herein, the first frame is configured to determine the concept event of the text data, and the second frame is configured to determine the concept event of the image data.

[ 00130]       Here, the lexicon refers to a library containing word meanings and is configured to determine the meaning of a word. In some embodiments, the lexicon may be WordNet. In the WordNet, nouns, adjectives and adverbs are stored in the database in the form of synsets, each synset represents a synset, and the synsets are in equilateral connection through a semantic relation and a lexical relation. The embodiments of this application do not limit the type of the lexicon. In some embodiments, the first frame may be Framenet and the second frame may be Imsitu, and the embodiments of this application do not limit the type of the first frame and the second frame.

[ 00131]       In some embodiments, a mapping relation is established between the lexicon and the first frame, a mapping relation is established between the lexicon and the second frame, and the lexicon, the first frame and the second frame are fused to obtain the multi-modal ontology.

[ 00132] In some embodiments, before the lexicon, the first frame and the second frame are fused, constructing the multi-modal ontology may also include the following operations.

[ 00133] The ontology is defined in the RDF by expanding the SEM as a basis for the knowledge representation.

[ 00134] In a case that the first frame is Framnet, the single event in the ontology may be expanded from the SEM and the FrameNet, that is, the RDF structure of the single event is expanded from the definitions of the SEM and the FrameNet. In some embodiments, the relation between the events includes the following three relations.

[ 00135] (1) The relation between the instance event and the concept event.

[ 00136] (2) The relation between the concept events, which is configured to represent the hierarchical relation between the concept events.

[ 00137] (3) Other various relations between the instance events, such as timing or causal relation and the like. In some embodiments, a link node may be added between the instance events to represent the relation between the instance events. Thus, in a case that a plurality of event relations exist between two instance events, the event relations may be quickly queried (only the required event relations need to be queried) and modified (expression of the edges between subject and object does not need to be modified, and only information of the link nodes may be modified), so that the ontology has good expandability.

[ 00138] In some embodiments, the ontology may also include a virtual node connected with the instance event to represent the source of this instance event. The edges of the virtual node may be configured to represent the trigger word, modality, sentence/image index and other information of the instance event respectively. The virtual node is a node designed for clear structure and needed for local structure integration, and the virtual node is queried only when needing to be queried, so that unnecessary query is reduced.

[ 00139] On the basis of constructing the multi-modal ontology, the implementation of S103 may include S1031 to S1032. At S1031, a first word sense matched with each

trigger word in the lexicon of the multi-modal ontology is determined. Here, the implementation of S1031 may determine the first word sense matched with each trigger word in the lexicon of the multi-modal ontology based on the word sense of the trigger word. For example, the trigger word is "buy," and then "buy" in the lexicon of the multi-modal ontology may be determined as the first word sense. At S1032, a first concept event matched with each first word sense is determined in a first frame of the multi-modal ontology. Here, due to the fact that the mapping relation is established between the first frame and the lexicon in the multi-modal ontology, after the first word sense in the lexicon is determined, the first concept event matched with each first word sense may be determined in the first frame.

[ 00140] In the embodiments of this application, the first word sense matched with each trigger word is determined in the lexicon of the multi-modal ontology, and then the first concept event matched with each first word sense is determined in the first frame of the multi-modal ontology, so that the concept event of at least one first instance event corresponding to the text data is determined, and the argument may be conveniently extracted from the text data subsequently.

[ 00141] At S104, a second concept event matched with each event type is determined based on the multi-modal ontology.

[ 00142] Here, on the basis of constructing the multi-modal ontology, as shown in FIG. 5B, the implementation of S104 may include S1041 to S1043 as follows.

[ 00143] At S1041, a first event matched with each event type is determined in a second frame of the multi-modal ontology.

[ 00144] Here, due to the fact that the event type of each event is carried in the second frame, the implementation of S1041 may determine the first event matched with each event type determined in S102 in the second frame based on the event type of each event carried in the second frame.

[ 00145] At S1042, a second word sense matched with each first event is determined in the lexicon of the multi-modal ontology.

- [ 00146]        Here, due to the fact that the mapping relation is established between the lexicon of the multi-modal ontology and the second frame, the second word sense matched with each first event may be determined in the lexicon of the multi-modal ontology.
- [ 00147]        At S1043, the second concept event matched with each second word sense is determined in the first frame of the multi-modal ontology.
- [ 00148]        Here, due to the fact that the mapping relation is established between the first frame and the lexicon in the multi-modal ontology, the first concept event matched with each second word sense may be determined in the first frame.
- [ 00149]        In the embodiments of this application, the first event matched with each event type is determined in the second frame of the multi-modal ontology, then the second word sense matched with each first event is determined in the lexicon of the multi-modal ontology, and finally the second concept event matched with each second word sense is determined in the first frame of the multi-modal ontology. Therefore, the concept event of at least one second instance event corresponding to the image data is determined, and the argument is conveniently extracted from the image data subsequently.
- [ 00150]        At S105, the MMEKG is constructed based on the first concept event matched with each trigger word and the second concept event matched with each event type.
- [ 00151]        After the concept event is obtained, the role matched with the concept event may be determined, so that the argument may be extracted from the text or the image according to the role, and the MMEKG is constructed.
- [ 00152]        In some embodiments, as shown in FIG. 5C, the implementation of S105 may include S1051 to S1053 as follows.
- [ 00153]        At S1051, event extraction is performed on the text data based on the first concept event matched with each trigger word to obtain an argument of each first instance event.
- [ 00154]        Here, the implementation of S1051 may include that: a role corresponding to

the first concept event is determined based on the first concept event; and then, based on the determined role, event extraction is performed on the text data to obtain the argument of each first instance event.

[ 00155]       At S1052, event extraction is performed on the image data based on the second concept event matched with each event type to obtain an argument of each second instance event.

[ 00156]       Here, the implementation of S1052 may include that: a role corresponding to the second concept event is determined based on the second concept event; and then, based on the determined role, event extraction is performed on the image data to obtain the argument of each second instance event.

[ 00157]       At 1053, the MMEKG is constructed based on the argument of each first instance event and the argument of each second instance event.

[ 00158]       In some embodiments, the MMEKG may include arguments, so that the MMEKG may be constructed from the argument of each first instance event and the argument of each second instance event.

[ 00159]       In the embodiments of this application, the first concept event of the text data and the second concept event of the image data are determined based on the multi-modal ontology, and then the MMEKG is constructed based on the first concept event and the second concept event, so that complementation and fusion among multi-modal information are realized, more information is provided for construction of the KG, and more possibilities are provided for knowledge mining.

[ 00160]       In some embodiments, after S105, as shown in FIG. 5D, the method further includes S106a, S107a and 108a.

[ 00161]       At S106a, at least one first instance event and at least one second instance event are determined as a first instance event set.

[ 00162]       The first instance event set includes the first instance event and the second instance event.

[ 00163]        At S107a, an event relation between two instance events in the first instance event set is determined.

[ 00164]        Here, the event relation between two instance events may include temporal and causal relations, semantic relations, co-occurrence relations.

[ 00165]        For the temporal and causal relations, the embodiments of this application provide a new method, that is, for constructing a document-level graph to infer the relation between the events globally. In the absence of clear temporal/causal indicators and complex heuristic rules, the method in the embodiments of this application may perform cross-sentence reasoning, which enables the embodiments of this application to simultaneously and effectively identify all temporal and causal relations of the document. In the embodiments of this application, the temporal and causal relations are jointly predicted into multi-label and multi-task classification, and a model is trained based on a causal time library and includes six relation types: before, after, during, including, included and causal relation.

[ 00166]        The co-occurrence relations include text co-occurrence and cross-modal co-occurrence. Herein, for text co-occurrence, if a conjunctive relation exists between the trigger words, text co-occurrence is considered. In the embodiments of this application, the conjunctive relation is identified through dependency parsing. For cross-modal co-occurrence, in the embodiments of this application, the events are extracted from the paired image characters respectively, and it is assumed that the events co-occur, so that mining of information between the paired image characters is achieved.

[ 00167]        For the semantic relations, the embodiments of this application provide a method for determining the semantic relation, correspondingly, the two instance events in S107a include a third instance event and a fourth instance event, as shown in FIG. 5E, and the implementation of S107a includes the following S1071a to S1073a.

[ 00168]        At S1071a, a grammatical relation between a trigger word of the third instance event and a trigger word of the fourth instance event is determined.

[ 00169]        Here, the grammatical relation includes csubj (subject clause), xcomp (open complement clause), ccomp (complement clause) or advcl (adverbial clause).

[ 00170]       At S1072a, a subordinate relation between the trigger word of the third instance event and an argument of the fourth instance event is determined.

[ 00171]       In some embodiments, S1072a may be implemented by judging whether the trigger word of the third instance event is the argument of the fourth instance event, and in the case of yes, the subordinate relation between the third instance event trigger word and the fourth instance event argument may be determined.

[ 00172]       At S1073a, when the grammatical relation meets a preset condition and the trigger word of the third instance event is the argument of the fourth instance event, the third instance event and the fourth instance event are in a semantic relation.

[ 00173]       Here, the preset condition is that in dependency parsing, the trigger words are connected through csubj (subject clause), xcomp (open complement clause), ccomp (complement clause) or advcl (adverbial clause).

[ 00174]       That is, the trigger word of the third instance event and the trigger word of the fourth instance event are connected through csubj (subject clause), xcomp (open complement clause), ccomp (complement clause) or advcl (adverbial clause). In a case that the trigger word of the third instance event is the argument of the fourth instance event, the third instance event and the fourth instance event are in a semantic relation. For example, two events are extracted from this sentence "eating too much fried chicken causes overweight": causing overweight (event A) and eating too much chicken (event B). Due to the fact that A is also an argument of the role influencing\_entity of B, the trigger word of the event A is "causing", the trigger word of the event B is "eating", the semantic relation between the trigger word of the event A and the trigger word of the event B is a subject clause, so that the event A and the event B are in a semantic relation.

[ 00175]       The embodiments of this application provide a new relation type (semantic relation) between the instance events, and provide a method for determining the semantic relation, so that the composition of the KG is enriched, the scale of the KG is enlarged, and more information is provided for mining.

[ 00176]       At S108a, the event relation between two instance events is added into the MMEKG.

[ 00177] In the embodiments of this application, the event relation between the two instance events is determined and added into the MMEKG, so that the composition of the KG is enriched, and more information is provided for mining

[ 00178] In some embodiments, after S105, the embodiments of this application also provides an ontology updating method, as shown in FIG. 5F, the method further includes S106b, S107b and S108b.

[ 00179] At S106b, a first significance score of the event relation between the two instance events in the first instance event set is determined.

[ 00180] Here, the event relation between the two instance events may include temporal and causal relations, semantic relations and co-occurrence relations. The calculation of the significance score considers the following two aspects: (1) the confidence score of the relation  $r$  between the instance events, and (2) the times of occurrence of a tail event (as little as possible), so that more important and meaningful information may be extracted.

[ 00181] At S107b, the relation between the concept events corresponding to the instance events is obtained by reasoning based on the first significance score.

[ 00182] Here, the implementation of S107b may include that: in a case that the first significance score exceeds a threshold, the relation between the instance events also exist in the concept events corresponding to the instance events through reasoning.

[ 00183] At S108b, the ontology is updated based on the relation, obtained by reasoning, between the concept events corresponding to the instance events.

[ 00184] Here, the implementation of S108b includes that: the relation between the concept events is added into the ontology, and the ontology is updated.

[ 00185] In the embodiments of this application, the first significance score of the event relation between the two instance events in the first instance event set is determined, then the relation between the concept events corresponding to the instance events is obtained by reasoning based on the first significance score, and then the ontology is updated, so that the hierarchical relation between the ontologies is more granular, and richer

knowledge information may be obtained conveniently.

[ 00186]       The embodiments of this application provide a multi-modal ontology construction method. As shown in FIG. 5G, the method includes S31 to S33 as follows.

[ 00187]       At S31, a word sense in a lexicon is matched with each event in a first frame, and a first mapping relation between the lexicon and the first frame is established.

[ 00188]       Herein, the first frame is configured to determine a concept event of text data. The first frame may be Framenet, and the lexicon may be WordNet.

[ 00189]       In some embodiments, the implementation of S31 may include S311 to S314 as follows.

[ 00190]       At S311, the event type of each event in the first frame is determined.

[ 00191]       Here, the event type of each event in the first frame may be an artificially labeled event type. In a case that the first frame is Framenet, each event in the Framenet carries its own event type.

[ 00192]       At S312, the cosine similarity between each word sense in the lexicon and the event type of each event in the first frame is determined.

[ 00193]       Here, the implementation of S312 may convert each word sense in the lexicon and the event type of each event in the first frame into a vector form (that is, by Sentence-BERT conversion vector form), and then the cosine similarity between each word sense in the lexicon and the event type of each event in the first frame is calculated.

[ 00194]       At S313, an event in the first frame matched with each word sense in the lexicon is determined based on the cosine similarity.

[ 00195]       Here, implementation of S313 may select the event corresponding to the event type with the highest cosine similarity as the event in the first frame matched with the word sense based on each word sense in the lexicon, thereby determining the event in the first frame matched with each word sense in the lexicon.

[ 00196]       At S314, the first mapping relation between the lexicon and the first frame is

established based on the event in the first frame matched with each word sense in the lexicon.

[ 00197]        Due to wide coverage of the event in the first frame FrameNet, for example, one chemical substance is converted to another is called Cause\_change, changing name is also called Cause\_change, while the WordNet has richer word sense, so that the WordNet and the FrameNet are aligned to make the concept event corresponding to the FrameNet more granular. In addition, the WordNet also brings hierarchical information to the FrameNet, so that the constructed KG information is richer.

[ 00198]        At S32, the word sense in the lexicon is matched with each event in a second frame, and a second mapping relation between the lexicon and the second frame is established

[ 00199]        Herein, the second frame is configured to determine a concept event of image data. The second frame may be Imsitu. The implementation of S32 may include S321 to S323.

[ 00200]        At S321, the event type of each event in the second frame is determined.

[ 00201]        Here, the event type of each event in the second frame may be an artificially labeled event type. In a case that the second frame is Imsitu, each event in the Imsitu carries its own event type.

[ 00202]        At S322, the word sense in the lexicon matched with each event in the second frame is determined based on the event type of each event in the second frame.

[ 00203]        Here, the implementation of S322 may determine a matching relation between each event in the second frame and the word sense in the lexicon by way of manual alignment according to the event type of each event in the second frame and the word sense in the lexicon.

[ 00204]        At S323, the second mapping relation between the lexicon and the second frame is established based on the word sense in the lexicon matched with each event in the second frame.

[ 00205]       Due to the fact that the second frame Imsitu is configured to describe the image event, an image event triggering event is triggered by a verb, and the verb corresponding to the event type in the Imsitu has no meaning, the verb needs to be aligned with the word sense under the WordNet. Due to the mapping relation between the Wordnet and the first frame Framenet, the event type under the FrameNet may be associated through the Wordnet, so that the concept event corresponding to the image is obtained. The Framenet is configured to obtain the concept event corresponding to the text, so that the multi-modal KG is constructed under the same ontology.

[ 00206]       At S33, the lexicon, the first frame and the second frame are fused based on the first mapping relation and the second mapping relation to obtain the multi-modal ontology.

[ 00207]       The second mapping relation exists between the second frame and the lexicon, and the first mapping relation exists between the first frame and the lexicon, so that the second frame and the first frame are associated through the lexicon, and the construction of the KG in two modes of text and image is realized.

[ 00208]       In the embodiment of this application, in a first aspect, due to the fact that the second frame Imsitu is configured to describe the image event, the image event triggering event is triggered by the verb, and the verb corresponding to the event type in the Imsitu has no meaning, the verb needs to be aligned with the word sense under the WordNet. Due to the mapping relation between the Wordnet and the first frame Framenet, the event type under the FrameNet may be associated through the Wordnet, so that the concept event corresponding to the image is obtained. The Framenet is configured to obtain the concept event corresponding to the text, so that the multi-modal KG is constructed under the same ontology. In a second aspect, due to wide coverage of the event in the first frame FrameNet and richer word sense of the lexicon WordNet, the WordNet and the FrameNet are aligned to make the concept event corresponding to the FrameNet more granular. In addition, the WordNet also brings hierarchical information to the FrameNet, so that the constructed KG information is richer.

[ 00209]       In some embodiments, after the multi-modal ontology is obtained at S33, the embodiments of this application also provide a multi-modal ontology updating method.

As shown in FIG. 5H, the method further includes S34 to S38.

[ 00210]       At S34, the concept event in the multi-modal ontology, a specific role in the concept event and a second instance event set matched with the concept event are acquired.

[ 00211]       Here, the concept event and the role in the concept event may be any concept event and any role. For example, the concept event is fry, the specific role is food. The second instance event set is all the instance events under the concept event fry.

[ 00212]       At S35, an argument set corresponding to the specific role in the second instance event set is determined.

[ 00213]       Here, continuing to illustrate by taking the concept event fry and the specific role food as an example, the arguments of the role food in the second instance event set include fish, a fish, the fish, a big fish, chicken and the chicken, which form the argument set.

[ 00214]       At S36, clustering processing is performed on the argument set to obtain a second significance score of each cluster.

[ 00215]       Here, the implementation of S36 may include that: word form reduction and heuristic clustering are performed on head words of all the arguments in the argument set, and then the second significance score of each type is calculated.

[ 00216]       Continuing to illustrate through the above example, fish, a fish, the fish and a big fish are all fish after the word form reduction, chicken and the chicken are all chicken after the word form reduction, and two types of fish and chicken are obtained after heuristic clustering is performed. The significance score of each type is calculated by considering the following two aspects: (1) a confidence score (the times of occurrence) of each type, (2) importance of each type, for example, the type is "I", then the importance is not high, and the weight of the type may be reduced.

[ 00217]       At S37, a new concept event is obtained by reasoning based on the second significance score of each type and the trigger word of the concept event.

- [ 00218]        Here, the implementation of S37 may include that: K types with the highest significance scores are selected and a new concept event is obtained by combining the role r and these names with the trigger word for reasoning. For example, the times of occurrence of fish is higher, the significance score is high, when K is 1, fish is combined with the trigger word fry and the role food, and a new concept event "fry fish" is derived.
- [ 00219]        At S38, the multi-modal ontology is updated based on the new concept event.
- [ 00220]        Here, the implementation of S38 may add the new concept event into the multi-modal ontology, so that the multi-modal ontology is updated.
- [ 00221]        In the embodiment of this application, the significance score of the argument in the instance event is determined, then the new conceptual event is obtained by reasoning in combination with the trigger word, and then the new concept event obtained by reasoning is added into the multi-modal ontology to update the multi-modal ontology, so that the relation in the ontology is richer and more granular, and the construction of a large-scale event KG is realized.
- [ 00222]        The PAIE provided in the embodiments of this application includes an event extraction method. The event extraction method is characterized by including the following operations that: an input text is acquired; herein, the input text includes a trigger word and at least one event argument; a first text feature vector set corresponding to the input text is extracted through an encoder and a decoder; a target extraction template corresponding to the trigger word is acquired; the target extraction template and the input text are decoded through the decoder to obtain an argument feature vector of each event argument; and a value range corresponding to each event argument is determined based on the first text feature vector set and the argument feature vector of each event argument.
- [ 00223]        In some embodiments, the method further includes an operation that: the input text and the target extraction template are converted into a text word vector set and a template word vector set. The operation that the first text feature vector set corresponding to the input text is extracted through the encoder and the decoder includes the following operations that: the text word vector set is encoded through the encoder to obtain a

second text feature vector set; and the text word vector set and the second text feature vector set are decoded through the decoder to obtain the first text feature vector set corresponding to the input text. Correspondingly, the operation that the target extraction template and the input text are decoded through the decoder to obtain the argument feature vector of each event argument includes an operation that: the template word vector set and the second text feature vector set are decoded through the decoder to obtain the argument feature vector of each event argument.

[ 00224] In some embodiments, the operation that the template word vector set and the second text feature vector set are decoded through the decoder to obtain the argument feature vector of each event argument includes the following operations that: the template word vector set and the second text feature vector set are decoded through the decoder to obtain a template feature vector set fused with input text information; and the argument feature vector of each event argument corresponding to each argument role is determined from the template feature vector set based on the position information of each argument role in the template word vector set.

[ 00225] In some embodiments, the operation that the argument feature vector of each event argument corresponding to each argument role is determined from the template feature vector set based on the position information of each argument role in the template word vector set includes the following operations that: for the position information of each argument role in the template word vector set, a feature vector of each event argument corresponding to the argument role is extracted from the template feature vector set; and average pooling is performed on the feature vector of each event argument to obtain the argument feature vector of the event argument.

[ 00226] Events are basic units that describe the activities and states of the objective world. A KG taking the events as nodes summarizes the events contained in each field in real life and the interactive relationship among the events, which may be very important in application of a series of downstream tasks such as common sense reasoning and the like. In recent years, there are a lot of related works on constructing event KGs: ATOMIC constructs a small-scale and high-quality event KG through manual tagging, and ASER constructs a large-scale and relatively-low-quality event KG through automatic rules. It is

still an unsolved problem to construct high-quality and relatively-large-scale event KGs without relying too much on manual tagging.

[ 00227] In addition, there is no strict design of a set of ontology for the event KG and consideration of complementation and fusion of multi-modal information (for example, text and visual images). Based on this, according to the embodiments of this application, by means of an event knowledge extraction system based on deep learning, more event knowledge is extracted at a relatively low cost (not relying on a large amount of manual tagging), fusion and complementation of the event knowledge between multi-modal information (images and text) are considered, and a set of high-quality and large-scale ontology with the objective use value is defined for the event KG.

[ 00228] The embodiments of this application provide a set of multi-modal event knowledge extraction system. The extraction system takes the forefront deep learning technology as the core, takes mature rules and templates as the assistance, and may quickly and efficiently extract a large amount of event knowledge from the text and images. The extraction system may achieve the best effect in the systems of the same type currently known.

[ 00229] The embodiments of this application provide a set of ontology of the MMEKG, define four different types of nodes and connecting edges for the graph, and stipulate association limits among different events and relations. By aligning external resources (fusing similar nodes in FrameNet and WordNet), a set of initialized ontology is obtained, and on the basis of the ontology, a set of hierarchical event/event relation ontology is induced by combining the extracted event instances.

[ 00230] The method provided in the embodiments of this application mainly includes the following two processes that: the first process is an ontology construction process, and the second process is a knowledge extraction process of the multi-modal event.

[ 00231] First process: Ontology construction process

[ 00232] A set of new ontology is constructed based on a conventional event KG ontology SEM, as shown in FIG. 3.

[ 00233] In FIG. 3, a dashed box represents a character attribute, and a solid box represents an entity/relation/event and the like. The entity is a basic unit in the KG that may have a relation with other entities or character attributes. A non-entity (character attribute) is a basic unit in the KG that may only exist as an attribute of a certain entity. For example, two triples (Obama, birth, USA) and (USA, establishment time, 1776) exist in the KG. (1) Obama and USA are entities, (2) birth and establishment time are in a relation, (3) 1776 is a non-entity/attribute.

[ 00234] The upper left part (solid boxes of grids, such as cpt\_Event\_565 and ins\_Event\_317) is the event-related ontology, the right part (grey solid box parts, such as [EventRel\_id] and [RelType\_id]) is the relation-related ontology, and the lower part (a part connected by a circle, that is, a virtual node, such as [span]) is the raw event information obtained by the extraction system. ekg represents a new type as defined by embodiments of this application, rdf represents a triple, ins\_Event represents an instance event, and cpt\_Event represents a concept event.

[ 00235] In general, the ontology of the entire event may be divided into two parts, a concept-level and an instance-level. For example, cpt\_Event\_565 and above belong to the concept-level, and ins\_Event\_317 and below belong to the instance-level. Firstly, the FrameNet and the WordNet are aligned to obtain an initial ontology. On the basis of the initial ontology, the instance events obtained by the event extraction system are counted, and the events with higher frequency of occurrence are merged into the concept-level to serve as a new ontology event. In addition, a similar method is adopted to merge the relation of the instance-level with higher frequency of occurrence into the concept-level.

[ 00236] Second process: Knowledge extraction process of the multi-modal event

[ 00237] The process is divided into three parts: a text knowledge extraction process, a visual knowledge extraction process and an event relation extraction process.

[ 00238] Text knowledge extraction includes the following operations that: firstly, POS tagging and Dependence parsing are performed on processed text; and the trigger word of the triggering event is obtained based on the specific rule. Herein, POS tagging may perform tagging by adopting a machine, and the specific rule may include that most of

verbs (a stop word table is maintained) and adjectives satisfying a plurality of dependencies. Then, WSD is performed on the vocabulary to obtain a basic event type. Finally, the event type is given, EAE is performed by adopting a prompt-based deep learning model, and a complete instance event, namely the PAIE method provided in the embodiments of this application, is obtained.

[ 00239]        Herein, POS tagging is configured to judge the part-of-speech of a word (verb, noun or adjective), and WSD is configured to determine the meaning of a polysemous word in the text (for example, the English word fire refers to termination of employment or firing). Stop Words are generally regarded in dictionaries as function words or non-search words. The stop word may also be understood as: in order to save storage space and improve search efficiency, certain words or phrases, called stop words, are automatically ignored in a search engine page or when processing a search request.

[ 00240]        Visual knowledge extraction is similar to text knowledge extraction, firstly, the event expressed by the image is determined through an Action Recognition technology, and then a related argument of the event is determined through a Situation Recognition technology.

[ 00241]        Event relation extraction is configured to classify the event relations into different types and perform differentiation processing. For the causal and temporal relations, event relation extraction is performed by adopting a relational graph reasoning technology based on Longformer. For the image-text co-occurrence relation, the relations in an image-text pair are aligned to obtain the corresponding relation. The events may be extracted from the images and text by utilizing paired data of the images and the text respectively. Two events extracted separately from a pair of image-text are considered to be aligned. For other relations, extraction is performed based on a defined rule. The classification of other event relations is complex, including text co-occurrence relations and hundreds of semantic relations.

[ 00242]        According to the technical solution, the embodiments of this application have the following characteristics that: 1) in the embodiments of this application, the automatic knowledge extraction system based on deep learning is adopted to extract the event knowledge to construct the KG, while in the related technology, manual tagging or

simple rule is adopted to extract the events to construct the KG. 2) The KG constructed by the embodiments of this application considers interaction and fusion of events with a plurality of modalities, while only single modality of the text is considered in the related technology. 3) in the embodiments of this application, an external database is fused with the extracted event instances, the event KG constructs a set of large-scale and high-quality ontology, while no ontology or only simple ontology exists in the related technology, and no new event ontology will be obtained based on the induction of the event instances. Therefore, the embodiments of this application have a faster and high-quality event knowledge extraction speed, and a more reasonable event knowledge organization and storage mode, and finally may realize interactive fusion of multi-modal knowledge. The methods and the systems provided in the embodiments of this application may be widely applied to products such as structured information extraction service products, service products related to common sense reasoning and the like.

[ 00243]       Based on the foregoing embodiments, the embodiments of this application provide an apparatus for KG construction. The apparatus includes the included modules and modules included in the units, and may be implemented by using a processor in a computer device. Definitely, the apparatus may alternatively be implemented by using a specific logic circuit. During implementation, the processor may be a Central Processing Unit (CPU), a Microprocessor Unit (MPU), a Digital Signal Processor (DSP), a Field Programmable Gate Array (FPGA) or the like.

[ 00244]       FIG. 6A is a schematic diagram of a composition structure of an apparatus for KG construction according to an embodiment of this application. As shown in FIG. 6A, the KG construction apparatus 610 includes a detection module 611, a classification module 612, a first determination module 613, a second determination module 614 and a construction module 615.

[ 00245]       The detection module 611 is configured to perform event detection on text data to obtain at least one trigger word of at least one first instance event corresponding to the text data.

[ 00246]       The classification module 612 is configured to perform image classification processing on image data to obtain at least one event type of at least one second instance

event corresponding to the image data.

[ 00247]       The first determination module 613 is configured to determine a first concept event matched with each trigger word based on a multi-modal ontology.

[ 00248]       The second determination module 614 is configured to determine a second concept event matched with each event type based on the multi-modal ontology.

[ 00249]       The construction module 615 is configured to construct a MMEKG based on the first concept event matched with each trigger word and the second concept event matched with each event type.

[ 00250]       In some embodiments, the multi-modal ontology includes a lexicon, a first frame and a second frame. Herein, the first frame is configured to determine the concept event of the text data, and the second frame is configured to determine the concept event of the image data. The first determination module includes a first determination unit and a second determination unit. The first determination unit is configured to determine a first word sense matched with each trigger word in the lexicon of the multi-modal ontology, and the second determination unit is configured to determine the first concept event matched with each first word sense in the first frame of the multi-modal ontology.

[ 00251]       In some embodiments, the second determination module includes a third determination unit, a fourth determination unit, and a fifth determination unit. The third determination unit is configured to determine a first event matched with each event type in the second frame of the multi-modal ontology. The fourth determination unit is configured to determine a second word sense matched with each first event in the lexicon of the multi-modal ontology. The fifth determination unit is configured to determine the second concept event matched with each second word sense in the first frame of the multi-modal ontology.

[ 00252]       In some embodiments, the construction module includes a first extraction unit, a second extraction unit, and a construction unit. The first extraction unit is configured to perform event extraction on the text data based on the first concept event matched with each trigger word to obtain an argument of each first instance event. The second extraction unit is configured to perform event extraction on the image data based on the

second concept event matched with each event type to obtain an argument of each second instance event. The construction unit is configured to construct the MMEKG based on the argument of each first instance event and the argument of each second instance event.

[ 00253] In some embodiments, the apparatus further includes: a third determination module, a fourth determination module, and an adding module. The third determination module is configured to determine at least one first instance event and at least one second instance event as a first instance event set. The fourth determination module is configured to determine an event relation between two instance events in the first instance event set. The adding module is configured to add the event relation between two instance events into the MMEKG.

[ 00254] In some embodiments, the two instance events include a third instance event and a fourth instance event. The fourth determination module is configured to determine a grammatical relation between a trigger word of the third instance event and a trigger word of the fourth instance event, and determine a subordinate relation between the trigger word of the third instance event and an argument of the fourth instance event. When the grammatical relation meets a preset condition and the trigger word of the third instance event is the argument of the fourth instance event, the third instance event and the fourth instance event are in a semantic relation.

[ 00255] In some embodiments, the apparatus further includes: a fifth determination module, a first reasoning module, and an updating module. The fifth determination module is configured to determine a first significance score of the event relation between two instance events in the first instance event set. The first reasoning module is configured to obtain the relation between the concept events corresponding to the instance events by reasoning based on the first significance score. The updating module is configured to update the ontology based on the relation, obtained by reasoning, between the concept events corresponding to the instance events.

[ 00256] Based on the foregoing embodiments, the embodiments of this application provide an apparatus for multi-modal ontology construction. The apparatus includes the included units and modules included in the units, and may be implemented by using a processor in a computer device. Definitely, the apparatus may alternatively be

implemented by using a specific logic circuit. During implementation, the processor may be a CPU, a MPU, a DSP, or a FPGA.

[ 00257] FIG. 6B is a schematic diagram of a composition structure of an apparatus for multi-modal ontology construction according to an embodiment of this application. As shown in FIG. 6B, the multi-modal ontology construction apparatus 620 includes a first establishment module 621, a second establishment module 622 and a fusion module 623.

[ 00258] The first establishment module 621 is configured to match at least one word sense in a lexicon with each event in a first frame respectively, and establish a first mapping relation between the lexicon and the first frame. The first frame is configured to determine a concept event of text data.

[ 00259] The second establishment module 622 is configured to match the at least one word sense in the lexicon with each event in a second frame respectively, and establish a second mapping relation between the lexicon and the second frame. The second frame is configured to determine a concept event of image data.

[ 00260] The fusion module 623 is configured to fuse the lexicon, the first frame and the second frame based on the first mapping relation and the second mapping relation to obtain the multi-modal ontology.

[ 00261] In some embodiments, the first establishment module includes: a sixth determination unit, a seventh determination unit, an eighth unit, and a first establishment unit. The sixth determination unit is configured to determine the event type of each event in the first frame. The seventh determination unit is configured to determine the cosine similarity between each word sense in the lexicon and the event type of each event in the first frame. The eighth unit is configured to determine an event in the first frame matched with each word sense in the lexicon based on the cosine similarity. The first establishment unit is configured to establish the first mapping relation between the lexicon and the first frame based on the event in the first frame matched with each word sense in the lexicon.

[ 00262] In some embodiments, the second establishment module includes: a ninth determination unit, a tenth determination unit and a second establishment unit. The ninth

determination unit is configured to determine the event type of each event in the second frame. The tenth determination unit is configured to determine the word sense in the lexicon matched with each event in the second frame based on the event type of each event in the second frame. The second establishment unit is configured to establish the second mapping relation between the lexicon and the second frame based on the word sense in the lexicon matched with each event in the second frame.

[ 00263] In some embodiments, the apparatus further includes: an acquisition module, a sixth determination module, a clustering module, a second reasoning module, and an updating module. The acquisition module is configured to acquire the concept event in the multi-modal ontology, a specific role in the concept event and a second instance event set matched with the concept event. The sixth determination module is configured to determine an argument set corresponding to the specific role in the second instance event set. The clustering module is configured to perform clustering processing on the argument set to obtain a second significance score of each type. The second reasoning module is configured to obtain a new concept event by reasoning based on the second significance score of each type and the trigger word of the concept event. The updating module is configured to update the multi-modal ontology based on the new concept event.

[ 00264] The descriptions of the above apparatus embodiments are similar to the descriptions of the above method embodiments, and have similar beneficial effects to the method embodiments. In some embodiments, functions or modules of the apparatus provided in the embodiments of the application may be configured to execute the method described in the above method embodiment. For technical details not disclosed in the apparatus embodiments of this application, refer to the descriptions of the method embodiments of this application for understanding.

[ 00265] It is to be noted that, in the embodiments of this application, if the above KG construction method or multi-modal ontology construction method is implemented in the form of a software functional module and sold or used as an independent product, it can be stored in the computer-readable storage medium. Based on such an understanding, the technical solutions of the embodiments of this application substantially or parts making contributions to the related art may be embodied in form of a software product, and the

software product is stored in a storage medium, including a plurality of instructions for causing a computer device (which may be a personal computer, a server, a network device or the like) to execute all or part of the method in each embodiment of the present disclosure. The foregoing storage medium includes: a USB flash disk, a mobile Hard Disk Drive (HDD), a Read-Only Memory (ROM), and various media that can store program codes, such as a magnetic disk, or an optical disk. In this way, the embodiments of this application are not limited to any specific hardware, software, or firmware, or any combination of hardware, software, and firmware.

[ 00266]       The embodiments of this application provide a computer device. The computer device includes a memory and a processor. The memory stores a computer program executable on the processor. The processor, when executing the computer program, implements the above method.

[ 00267]       The embodiments of this application provide a computer-readable storage medium. The computer-readable storage medium stores a computer program. The above method is implemented when the computer program is performed by a processor. The computer-readable storage medium may be transitory or non-transitory.

[ 00268]       The embodiments of this application provide a computer program. The computer program includes computer-readable codes. When the computer-readable codes are operated in a computer device, a processor in the computer device executes and implements part or all of steps is the above method.

[ 00269]       The embodiments of this application provide a computer program product. The computer program product includes a non-transitory computer-readable storage medium storing a computer program. When the computer program is read and executed by a computer, part or all of steps in the above method are implemented. The computer program product can be specifically implemented in a manner of hardware, software or a combination thereof. In some embodiments, the computer program product is embodied as a computer storage medium. In some other embodiments, the computer program product is embodied as a software product, such as a Software Development Kit (SDK) and the like.

[ 00270]        If the technical solution of this application relates to personal information, before the personal information is processed, the product applying the technical solution of this application clearly informs the personal information processing rule and obtains individual consent. If the technical solution of this application relates to sensitive personal information, before the sensitive personal information is processed, the product applying the technical solution of this application obtains individual consent, and meets the requirement of "express consent" at the same time. For example, a clear and obvious identifier is set at a personal information acquisition apparatus such as a camera to inform that the personal information is about to be acquired when entering a personal information acquisition range, and if an individual voluntarily enters the acquisition range, the individual is regarded as agreeing to acquire the personal information; or on a personal information processing device, in a case that a personal information processing rule is informed by using an obvious identifier/information, personal authorization is obtained by means of pop-up window information or asking a person to upload personal information of the person by himself or the like. Herein, the personal information processing rule may include information such as a personal information processing person, a personal information processing purpose, a processing mode, a processed personal information type and the like.

[ 00271]        Herein, it is to be noted that, the above description of the various embodiments tends to emphasize the differences between the various embodiments, and the identicalness or similarities can be referred to each other. The descriptions of the above device, storage medium, computer program, and computer program product embodiments are similar to the descriptions of the above method embodiments, and have similar beneficial effects to the method embodiments. For technical details not disclosed in the embodiments of the device, storage medium, computer program and computer program product of this application, refer to the description of the method embodiments of this application for understanding.

[ 00272]        It is to be noted that, FIG. 7 is a schematic diagram of a hardware entity of a computer device according to an embodiment of this application. As shown in FIG. 7, a hardware entity of the computer device 700 includes a processor 701, a communication interface 702, and a memory 703.

[ 00273]       The processor 701 usually controls an overall operation of the computer device 700.

[ 00274]       The communication interface 702 may cause the computer device to achieve communication with other terminals or servers through a network.

[ 00275]       The memory 703 is configured to store an instruction and an application executable by the processor 701, or may cache data to be processed or processed by the processor 701 and each module in the computer device 700 (such as, image data, audio data, voice communication data, and video communication data), which may be implemented by using FLASH or a Random Access Memory (RAM). Data transmission among the processor 701, the communication interface 702, and the memory 703 may be achieved by using a bus 704.

[ 00276]       It is to be understood that "one embodiment" and "an embodiment" mentioned in the whole specification mean that specific features, structures or characteristics related to the embodiment is included in at least one embodiment of this application. Therefore, "in one embodiment" or "in an embodiment" appearing at any place of the whole specification does not always refer to the same embodiment. In addition, these specific features, structures or characteristics may be combined in one or more embodiments in any proper manner. It is to be understood that, in various embodiments of this application, the sequence number of each step/process does not mean the sequence of execution. The execution sequence of each step/process should be determined by its functions and internal logic, which should not constitute any limitation on the implementation process of the embodiments of this application. The serial numbers of the foregoing embodiments of this application are merely for description, and do not represent the superiority or inferiority of the embodiments.

[ 00277]       It is to be noted that terms "include" and "comprise" or any other variant thereof is intended to cover nonexclusive inclusions herein, so that a process, method, object or apparatus including a series of components not only includes those components but also includes other components which are not clearly listed or further includes components intrinsic to the process, the method, the object or the apparatus. Under the condition of no more limitations, a component defined by the statement "including

a/an....." does not exclude existence of the same other components in a process, method, object or apparatus including the component.

[ 00278] In several embodiments provided by this application, it is to be understood that the disclosed device and method may be implemented in other ways. The device embodiment described above is only schematic, and for example, division of the units is only logic function division, and other division manners may be adopted during practical implementation. For example, a plurality of units or components may be combined or integrated into another system, or some characteristics may be neglected or not executed. In addition, coupling or direct coupling or communication connection between each displayed or discussed component may be indirect coupling or communication connection, implemented through some interfaces, of the device or the units, and may be electrical and mechanical or adopt other forms.

[ 00279] The units described as separate parts may or may not be physically separated, and parts displayed as units may or may not be physical units, and namely may be located in the same place, or may also be distributed to multiple network units. Part of all of the units may be selected according to a practical requirement to achieve the purposes of the solutions of the embodiments.

[ 00280] In addition, each functional unit in each embodiment of this application may be integrated into a processing unit, each unit may also serve as an independent unit and two or more than two units may also be integrated into a unit. The integrated unit may be implemented in a hardware form and may also be implemented in form of hardware and software functional unit.

[ 00281] Those of ordinary skill in the art should know that all or part of the steps of the method embodiment may be implemented by related hardware instructed through a program, the program may be stored in a computer-readable storage medium, and the program is executed to execute the steps of the method embodiment. The storage medium which may be a volatile storage medium includes: a mobile storage device, an ROM, and various media that can store program codes, such as a magnetic disk, or an optical disk.

[ 00282] Or, if the integrated unit is implemented in the form of the software functional

unit and sold or used as an independent product, it can be stored in the computer readable storage medium. Based on such an understanding, the technical solutions of this application substantially or parts making contributions to the related art may be embodied in form of a software product, and the computer software product is stored in a storage medium, including a plurality of instructions for causing a computer device (which may be a personal computer, a server, a network device or the like) to execute all or part of the method in each embodiment of this application. The foregoing storage medium includes a portable storage device, an ROM, and various media that can store program codes, such as a magnetic disk, or an optical disk.

[ 00283]       The above is only the implementations of this application and not intended to limit the scope of protection of this application. Any variations or replacements apparent to those skilled in the art within the technical scope disclosed by this application shall fall within the scope of protection of this application.

## CLAIMS

1. A method for knowledge graph construction, comprising:

performing event detection on text data to obtain at least one trigger word of at least one first instance event corresponding to the text data;

performing image classification processing on image data to obtain at least one event type of at least one second instance event corresponding to the image data;

determining a first concept event matched with each trigger word based on a multi-modal ontology;

determining a second concept event matched with each event type based on the multi-modal ontology; and

constructing a multi-modal event knowledge graph based on the first concept event matched with each trigger word and the second concept event matched with each event type.

2. The method of claim 1, wherein the multi-modal ontology comprises a lexicon, a first frame and a second frame, wherein the first frame is configured to determine the concept event of the text data, and the second frame is configured to determine the concept event of the image data;

determining the first concept event matched with each trigger word based on the multi-modal ontology comprises:

determining a first word sense matched with each trigger word in the lexicon of the multi-modal ontology; and

determining the first concept event matched with each first word sense in the first frame of the multi-modal ontology.

3. The method of claim 2, wherein determining the second concept event matched with each event type based on the multi-modal ontology comprises:

determining a first event matched with each event type in the second frame of the multi-modal ontology;

determining a second word sense matched with each first event in the lexicon of the multi-modal ontology; and

determining the second concept event matched with each second word sense in the first frame of the multi-modal ontology.

4. The method of any one of claims 1 to 3, wherein constructing the multi-modal event knowledge graph based on the first concept event matched with each trigger word and the second concept event matched with each event type comprises:

performing event extraction on the text data based on the first concept event matched with each trigger word to obtain an argument of each first instance event;

performing event extraction on the image data based on the second concept event matched with each event type to obtain an argument of each second instance event; and

constructing the multi-modal event knowledge graph based on the argument of each first instance event and the argument of each second instance event.

5. The method of any one of claims 1 to 4, further comprising:

determining the at least one first instance event and the at least one second instance event as a first instance event set;

determining an event relation between two instance events in the first instance event set; and

adding the event relation between the two instance events into the multi-modal event knowledge graph.

6. The method of claim 5, wherein the two instance events comprise a third instance event and

a fourth instance event; determining the event relation between the two instance events in the first instance event set comprises:

determining a grammatical relation between a trigger word of the third instance event and a trigger word of the fourth instance event; and

determining a subordinate relation between the trigger word of the third instance event and an argument of the fourth instance event;

when the grammatical relation meets a preset condition and the trigger word of the third instance event is an argument of the fourth instance event, the third instance event and the fourth instance event are in a semantic relation.

7. The method of claim 5 or 6, further comprising:

determining a first significance score of the event relation between the two instance events in the first instance event set;

obtaining a relation between concept events corresponding to the two instance events by reasoning based on the first significance score; and

updating the multi-modal ontology based on the relation, obtained by reasoning, between the concept events corresponding to the two instance events.

8. A method for multi-modal ontology construction, comprising:

matching at least one word sense in a lexicon with each event in a first frame respectively, and establishing a first mapping relation between the lexicon and the first frame, wherein the first frame is configured to determine a concept event of text data;

matching the at least one word sense in the lexicon with each event in a second frame respectively, and establishing a second mapping relation between the lexicon and the second frame, wherein the second frame is configured to determine a concept event of image data; and

fusing the lexicon, the first frame and the second frame based on the first mapping relation and the second mapping relation to obtain the multi-modal ontology.

9. The method of claim 8, wherein matching the at least one word sense in the lexicon with each event in the first frame respectively, and establishing the first mapping relation between the lexicon and the first frame comprises:

determining an event type of each event in the first frame;

determining a cosine similarity between each word sense in the lexicon and the event type of each event in the first frame;

determining an event in the first frame matched with each word sense in the lexicon based on the cosine similarity; and

establishing the first mapping relation between the lexicon and the first frame based on the event in the first frame matched with each word sense in the lexicon.

10. The method of claim 8 or 9, wherein matching at least one the word sense in the lexicon with each event in the second frame respectively, and establishing the second mapping relation between the lexicon and the second frame comprises:

determining an event type of each event in the second frame;

determining the word sense in the lexicon matched with each event in the second frame based on the even type of each event in the second frame; and

establishing the second mapping relation between the lexicon and the second frame based on the word sense in the lexicon matched with each event in the second frame.

11. The method of any one of claims 8 to 10, wherein after the multi-modal ontology is obtained, the method further comprises:

acquiring a concept event in the multi-modal ontology, a specific role in the concept event and a second instance event set matched with the concept event;

determining an argument set corresponding to the specific role in the second instance event set;

performing clustering processing on the argument set to obtain a second significance

score of each cluster;

obtaining a new concept event by reasoning based on the second significance score of each type and a trigger word of the concept event; and

updating the multi-modal ontology based on the new concept event.

12. An apparatus for knowledge graph construction, comprising:

a detection module, configured to perform event detection on text data to obtain at least one trigger word of at least one first instance event corresponding to the text data;

a classification module, configured to perform image classification processing on image data to obtain at least one event type of at least one second instance event corresponding to the image data;

a first determination module, configured to determine a first concept event matched with each trigger word based on a multi-modal ontology;

a second determination module, configured to determine a second concept event matched with each event type based on the multi-modal ontology; and

a construction module, configured to construct a multi-modal event knowledge graph based on the first concept event matched with each trigger word and the second concept event matched with each event type.

13. The apparatus of claim 12, wherein the multi-modal ontology includes a lexicon, a first frame and a second frame, wherein the first frame is configured to determine the concept event of the text data, and the second frame is configured to determine the concept event of the image data, wherein the first determination module comprises:

a first determination unit, configured to determine a first word sense matched with each trigger word in the lexicon of the multi-modal ontology; and

a second determination unit, configured to determine the first concept event matched

with each first word sense in the first frame of the multi-modal ontology.

14. The apparatus of claim 13, wherein the second determination module comprises:

a third determination unit, configured to determine a first event matched with each event type in the second frame of the multi-modal ontology;

a fourth determination unit, configured to determine a second word sense matched with each first event in the lexicon of the multi-modal ontology, and

a fifth determination unit, configured to determine the second concept event matched with each second word sense in the first frame of the multi-modal ontology.

15. An apparatus for multi-modal ontology construction, comprising:

a first establishment module, configured to match at least one word sense in a lexicon with each event in a first frame respectively, and establish a first mapping relation between the lexicon and the first frame, wherein the first frame is configured to determine a concept event of text data;

a second establishment module, configured to match the at least one word sense in the lexicon with each event in a second frame respectively, and establish a second mapping relation between the lexicon and the second frame, wherein the second frame is configured to determine a concept event of image data; and

a fusion module, configured to fuse the lexicon, the first frame and the second frame based on the first mapping relation and the second mapping relation to obtain the multi-modal ontology.

16. The apparatus of claim 15, wherein the first establishment module comprises:

a sixth determination unit, configured to determine the event type of each event in the first frame;

a seventh determination unit, configured to determine the cosine similarity between

each word sense in the lexicon and the event type of each event in the first frame;

an eighth unit, configured to determine an event in the first frame matched with each word sense in the lexicon based on the cosine similarity; and

a first establishment unit, configured to establish the first mapping relation between the lexicon and the first frame based on the event in the first frame matched with each word sense in the lexicon.

17. The apparatus of claim 15 or 16, wherein the second establishment module comprises:

a ninth determination unit, configured to determine the event type of each event in the second frame;

a tenth determination unit, configured to determine the word sense in the lexicon matched with each event in the second frame based on the event type of each event in the second frame; and

a second establishment unit, configured to establish the second mapping relation between the lexicon and the second frame based on the word sense in the lexicon matched with each event in the second frame.

18. A computer device, comprising a memory and a processing, wherein the memory stores a computer program executable on the processor, and the processor, when executing the program, implements steps in the method according to any of claims 1 to 11.

19. A computer-readable storage medium, having a computer program stored thereon, wherein steps in the method of any of claims 1 to 11 are implemented when the computer program is performed by a processor.

20. A computer program, comprising computer-readable codes, wherein when the computer-readable codes are operated in a computer device, a processor in the computer device implements steps in the method according to any of claims 1 to 11.