5-2022

MMEKG: Multi-modal Event Knowledge Graph towards universal representation across modalities

Yubo MA

Zehao WANG

Mukai LI

Yixin CAO
Singapore Management University, yxcao@smu.edu.sg

Meiqi CHEN

See next page for additional authors

Citation

MA, Yubo; WANG, Zehao; LI, Mukai; CAO, Yixin; CHEN, Meiqi; LI, Xinze; SUN, Wenqi; DENG, Kunquan; WANG, Kun; SUN, Aixin; and SHAO, Jing. MMEKG: Multi-modal Event Knowledge Graph towards universal representation across modalities. (2022). *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, Dublin, Ireland, 2022 May 22-27*. 231-239.
Available at: https://ink.library.smu.edu.sg/sis_research/7445

MMEKG: Multi-modal Event Knowledge Graph towards Universal Representation across Modalities

Yubo Ma[1†*], **Zehao Wang**[2†*], **Mukai Li**[3†*], **Yixin Cao**[4*], **Meiqi Chen**[5†*],
Xinze Li[1], **Wenqi Sun**[3], **Kunquan Deng**[3], **Kun Wang**[3], **Aixin Sun**[1], **Jing Shao**[3‡]
[1] S-Lab, Nanyang Technological University [2] KU Leuven [3] SenseTime Research
[4] Singapore Management University [5] Peking University
yubo001@e.ntu.edu.sg

Abstract

Events are fundamental building blocks of real-world happenings. In this paper, we present a large-scale, multi-modal event knowledge graph named MMEKG. MMEKG unifies different modalities of knowledge via events, which complement and disambiguate each other. Specifically, MMEKG incorporates (i) over 990 thousand concept events with 644 relation types to cover most types of happenings, and (ii) over 863 million instance events connected through 934 million relations, which provide rich contextual information in texts and/or images. To collect billion-scale instance events and relations among them, we additionally develop an efficient yet effective pipeline for textual/visual knowledge extraction system. We also develop an induction strategy to create million-scale concept events and a schema organizing all events and relations in MMEKG. To this end, we also provide a pipeline[1] enabling our system to seamlessly parse texts/images to event graphs and to retrieve multi-modal knowledge at both concept- and instance-levels.

1 Introduction

Recently, many Knowledge Graphs (KGs) have been curated (*e.g.,* Wikidata (Vrandečić and Krötzsch, 2014)) and successfully applied to various applications, ranging from information extraction (Lai et al., 2021) to information retrieval (Dong et al., 2014). KGs typically store billions of world facts in a directed graph, where nodes denote entities and edges denote their relations. Although simple yet effective, the expression ability of such entity-centric KGs is limited (Liu et al., 2020). How we can represent more complex knowledge, such as events, situations, or different modalities, becomes a key question for broader applications.

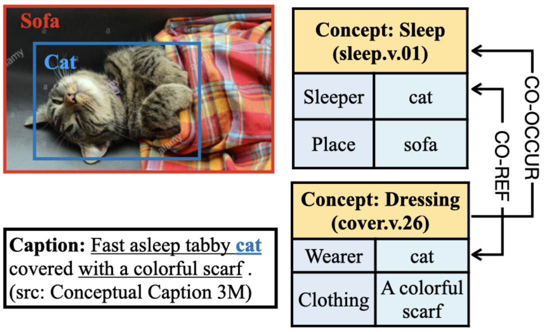

Sofa
Cat
Concept: Sleep (sleep.v.01)
Sleeper
cat
Place
sofa
Concept: Dressing (cover.v.26)
Wearer
cat
Clothing
A colorful scarf
CO-OCCUR
CO-REF
Caption: Fast asleep tabby cat covered with a colorful scarf .
(src: Conceptual Caption 3M)

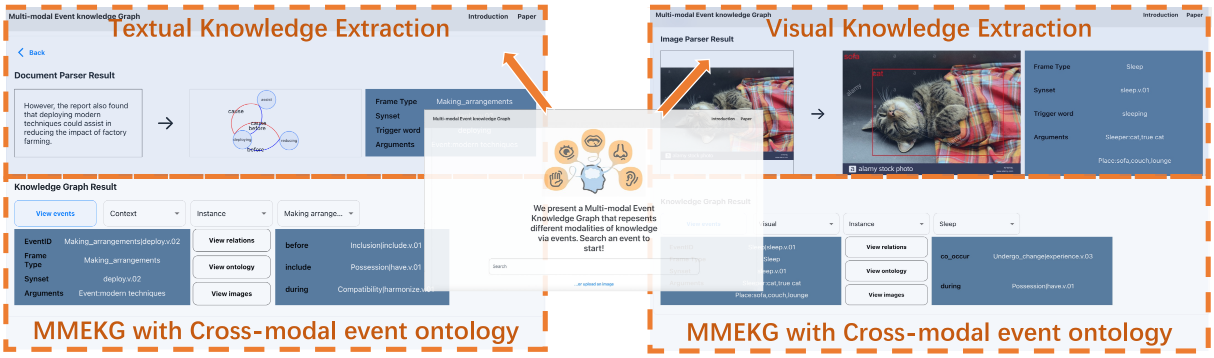
Figure 1: Examples of visual and textual events, and their relations. CO-REF denotes co-reference.

In this paper, we present a large-scale **M**ulti-**M**odal **E**vent **K**nowledge **G**raph (MMEKG) that bridges, complements, and disambiguates different modalities of knowledge, for better understanding or reasoning. Similar to real-world happenings, MMEKG takes events as its basic building blocks. Each event is defined by a concept, several arguments, and corresponding roles. Among events are various types of relations, such as causal, temporal, or sub-event relations. Thus eneities can be arguments in KGs. Figure 1 shows two example events: a visual *sleep* event with arguments *cat* (sleeper) and *sofa* (place), and a textual *dressing* event with arguments *cat* (wearer) and *scarf* (clothing), where argument roles are in brackets. The two events not only bridge the text and image with complementary arguments but also offer underlying commonsense knowledge — covering with a scarf usually happens when sleeping.

Compared with existing event KGs (Speer et al., 2016; Zhang et al., 2020; Hwang et al., 2021), MMEKG advances this field in the following three aspects: (1) A **large-scale ontology** contains 990 thousand concept events and 644 relation types, which covers most types of real-world happenings. (2) **Multi-modal knowledge** is naturally fused. To our best knowledge, it is the first event KG that bridges different modalities of data through fine-

[*] Equal Contribution.
[†] Work was done when Yubo, Zehao, Mukai and Meiqi were intern researchers at SenseTime Research.
[‡] Corresponding Author.
[1] System page: https:www.mmekg.com.

Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics
System Demonstrations, pages 231 - 239
May 22-27, 2022 ©2022 Association for Computational Linguistics



Document Parser Result

However, the report also found that deploying modern techniques could assist in reducing the impact of factory farming.

Frame Type Making_arrangements
Synset
Trigger word
Arguments

Knowledge Graph Result

Context
Instance
Making arrange...

EventID Making_arrangements|deploy.v.02
Frame Type Making_arrangements
Synset deploy.v.02
Arguments Event:modern techniques

before Inclusion|include.v.01
include Possession|have.v.01
during Compatibility|harmonize.v.01

We present a Multi-modal Event Knowledge Graph that repesents different modalities of knowledge via events. Search an event to start!

Image Parser Result

Frame Type Sleep
Synset sleep.v.01
Trigger word sleeping
Arguments Sleeper:cat,true cat
Place:sofa,couch,lounge

Knowledge Graph Result

Visual
Instance
Sleep

EventID Sleep|sleep.v.01
Frame Type Sleep
Synset sleep.v.01
Arguments Sleeper:cat,true cat
Place:sofa,couch,lounge

co_occur Undergo_change|experience.v.03
during Possession|have.v.01

Figure 2: Illustration of Demo System. Any input texts/images can be parsed into event graphs, where nodes denote instance events and edges denote event-event relations. Each instance event also refers to detailed information: concept event, synset, arguments with corresponding roles, and the linked neighbors in MMEKG (blue tables). Note that a single image mainly contains one event.

grained alignments of events and arguments. (3) The **integration of concept and instance events** not only makes it possible to enlarge the ontology from instance events but also provides concept-level commonsense knowledge with contextual instances for comprehensive reasoning.

There are mainly two steps to build MMEKG. (1) To construct a schema and acquire concept events, we first manually combine FrameNet (Baker et al., 1998) and WordNet (Fellbaum, 1998) to initialize a high-quality event ontology; we then expand it automatically via ontology induction from instance events. For flexibility and exchangeability, we extend the Simple Event Model (SEM) (Van Hage et al., 2011) to define our ontology in Resource Description Framework (RDF). (2) To extract instance events from either texts or images, we developed a knowledge extraction system to support fast and massive extraction under the practical scenario. This system consists of event extraction and event relation extraction in both modalities, as well as the alignment between them. In addition, this system can parse any input texts/images to event graphs and seamlessly retrieve multi-modal knowledge from MMEKG.

To cover a variety of events, we apply our extraction system into multiple sources, including C4 News[2], Wikipedia[3], Bookcorpus[4], and CC3M&12M (Sharma et al., 2018; Changpinyo et al., 2021). These data sources result in 863 million instance events and 934 million relations. To ensure its quality, we evaluate both our extraction system and the constructed MMEKG. Compared with state-of-the-art models of each sub-tasks, our methods achieve comparable or better performance on standard benchmarks. The adaptation to practical corpus led to no significant degradation. We sample thousands of events and relations from MMEKG for manual evaluation. The precision is acceptable at both concept and instance levels.

2 Overview of MMEKG

2.1 Definitions

Our proposed MMEKG, as shown in Figure 3, is different from traditional event-centric KGs and has four types of nodes and four types of relations. Nodes include concept events, instance events, entities, and non-entity arguments *e.g.,* literals. Among them, concept events (color in purple in Figure 3) are modality agnostic and provide high-level summarization of instance events (color in yellow), and entities/literals (color in blue) could be event arguments. The four types of relations contain (1) relation between instance events. Such type of relation can be further categorized into more fine-grained sub-types, such as temporal, causal, co-occur, and other semantic relations, (2) relation between concept events, named as *subclassOf* which denotes a hierarchical relation, (3) relation between concept events and instance events, named as *instanceOf* relation that integrates concept and instance events, and (4) role relations that reflect the roles of arguments (entities or non-entities) to the linked events. Different concept events have different roles. Formally, we have:

Definition 1 MMEKG $= \{(h,r,t)|h,t \in \mathcal{E}, r \in R\}$. $\mathcal{E} = \mathcal{E}_{cpt} \bigcup \mathcal{E}_{ins} \bigcup \mathcal{E}_{ent} \bigcup \mathcal{E}_{nent}$, where $\mathcal{E}_{cpt}$, $\mathcal{E}_{ins}$, $\mathcal{E}_{ent}$, and $\mathcal{E}_{nent}$ represent

[2] https://www.tensorflow.org/datasets/catalog/c4
[3] https://dumps.wikimedia.org/enwiki/
[4] https://www.gutenberg.org/

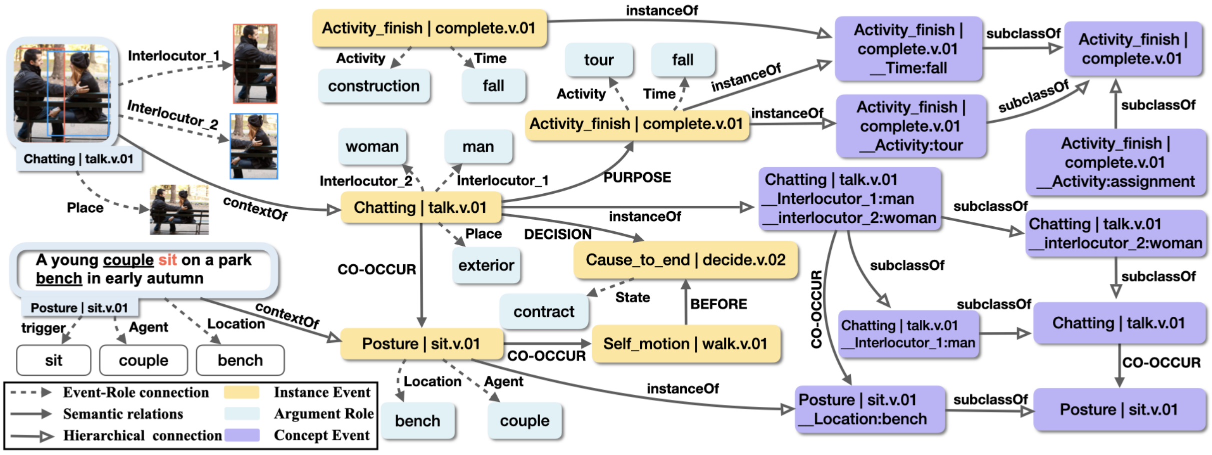


Figure 3: Three levels of MMEKG are illustrated from left to right. The left part is extracted multimodal context. The middle part shows the instance events aggregated from raw context. The right part are inducted concept events.

the set of concept event, instance event, entities, and non-entities, respectively. $\mathcal{R} = \mathcal{R}_{ins-ins} \bigcup \mathcal{R}_{cpt-cpt} \bigcup \mathcal{R}_{cpt-ins} \bigcup \mathcal{R}_{role}$, where $\mathcal{R}_{ins-ins}$ and $\mathcal{R}_{cpt-cpt}$ represent the set of relations between instance events or between concept events, $\mathcal{R}_{cpt-ins}$ represents the set of relations between instance events and concept events, and $\mathcal{R}_{role}$ denotes the set of argument roles. $w(h, r, t)$ denotes the relation weight of the triple (h, r, t) in MMEKG, i.e., the confidence score of being true.

2.2 User Interface and System Architecture

As shown in Figure 2, based on MMEKG and the extraction system, we have developed a prototype system that can parse arbitrary texts or images to an event graph, where the nodes denote instance events and the edges denote their relations. For each instance event, we link it to a concept event in MMEKG by identifying the trigger word and its synset (Event Detection). According to the concept event and corresponding roles, we also extract arguments, either a span in texts or a region in images (Argument Extraction). These modules consist of two main components: **Textual Knowledge Extraction** and **Visual Knowledge Extraction** (no trigger word). Another main component is **Event Relation Extraction** which extracts various relations among events, including the fusion of textual and visual events. Note that concept events, synsets, and relation types, are defined by our **cross-modal event ontology**. The linked neighbors in MMEKG are also shown below for better understanding. The detailed architectures behind the demo system, MMEKG and the extraction system, are shown in Figure 3 and Figure 5 respectively.

3 Cross-modal Event Ontology

Ontology is critical because it not only confines what types of knowledge are concerned but also offers a reasoning ability — only the induction from instances to concepts brings new knowledge, *i.e.,* from the special to the general. The deduction from concepts to instances has no uncertainty but provides additional information. In this section, we introduce our RDF Schema to model ontology data (Section 3.1), an initial ontology by combining external resources (Section 3.2), and ontology induction for continuous expansion (Section 3.3).

3.1 Schema

Following prior work (Gottschalk and Demidova, 2019), we inherit and extend the basic Simple Event Model (SEM) (Van Hage et al., 2011; McBride, 2004) as a knowledge representation basis. An example schema is shown in Figure 4.

Single event representation is extended from SEM and FrameNet. (1) Each role has an associated *ekg:[role]* connecting instance event $e \in \mathcal{E}_{ins}$ and argument $a \in \mathcal{E}_{ent} \bigcup \mathcal{E}_{nent}$. (2) We additionally add virtual nodes connecting instance events with edge *ekg:contextOf* to represent a source of such event. Edges from the virtual node like *ekg:trigger*, *ekg:modality* and *ekg:content* indicate the trigger word, modality and sentence/image index of this source respectively.

Event-event Relation mainly includes (1) *rdf:instanceOf* to integrate instance and concept events, (2) *rdf:subclassOf* that indicates the hierarchy of concept events, and (3) other relations among instance events, such as temporal or causal relations. For such relations, we design a link-

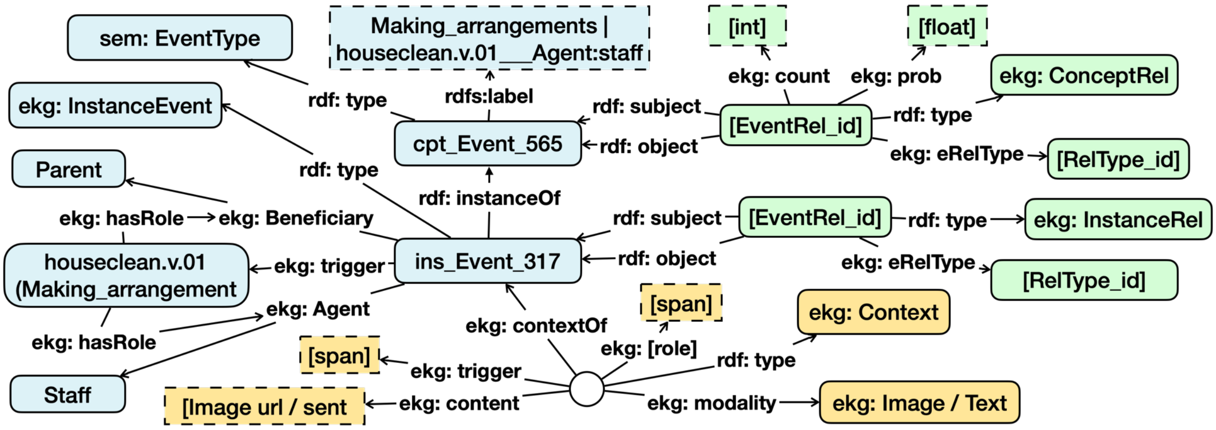


Figure 4: Illustration of Schema designed in MMEKG. Dashed boxes indicate literals and solid boxes indicate events, entities, and relations. We use different colors to represent different types of schema. **Blue**: Event-related. **Green**: Relation-related. **Yellow**: Information of related texts/images from which the system extracts instance events. The uncolored circle is a virtual node connecting an instance event and its source information.

ing node marked by [*EventRel_id*]. There are two advantages to this design: (a) Good extensibility. For possible N-to-1 event relations, multiple subjects and objects can be organized through the linking nodes. Conventional *subj-rel-obj* tuples cannot handle this case. (b) Integration of informative statistics and supplements. For example, we add the frequency and confidence score (obtained from ontology induction in Section 3.3) as prior to the events for reasoning with uncertainty.

3.2 Ontology Initialization

Based on schema, we initialize the ontology by merging WordNet (Fellbaum, 1998), FrameNet (Baker et al., 1998), and imSitu (Yatskar et al., 2016) Ontology. In specific, we map each verb and adjective synset in WordNet to a frame in FrameNet (for example, roast.v.01 ⟶ Apply_heat). The frames are high-level concept events, and the aligned synsets become fine-grained concept events. Moreover, the WordNet taxonomy brings hierarchical information. For mapping, we first jointly consider the result from structural mapping (Leseva and Stoyanova, 2019) and cosine-similarity score between definitions about synsets and frames given by Sentence-BERT (Reimers and Gurevych, 2019). We randomly sample 100 synset-frame pairs to check whether the definitions of mapped synset and frame align well, and find 89% pairs are reasonable. Then we extend the ontology from imSitu dataset by manually aligning WordNet synset to annotated frame as our visual ontology.

3.3 Ontology Induction

This section details how to expand the initial ontology from the perspectives of hierarchical taxonomy and relation types.

Taxonomy Induction finds more fine-grained concept events hierarchically. For example, both complete, complete a tour and complete a tour in fall belong to the initialized concept event Activity_finish:complete.v.01, while they represent events with different granularity. Therefore we hope to discriminate them with a more hierarchical and fine-grained taxonomy structure.

Given an initialized concept event o and one of its specific roles r, we first select all arguments connected by role r with an instance event categorized to o. Then we cluster these arguments heuristically by lemmatizing the headword of each phrase. We further name each cluster by that lemmatized headword and calculate a salience score for each cluster by jointly considering (1) the confidence score w of each event-role-argument triple clustered in and (2) how much information each cluster name provides. Finally, we select K clusters with the highest salience scores and create new concept events by combining role r and these names with their trigger words. Corresponding instance events are also categorized into these newly derived concept events. As shown in Figure 3, we derive new concept events such as complete.v.01__Activity:tour and complete.v.01__Activity:tour__Time:fall. These fine-grained concept events summarize instance events via *instanceOf* relations and are summarized by complete.v.01 with *subclassOf* relations.

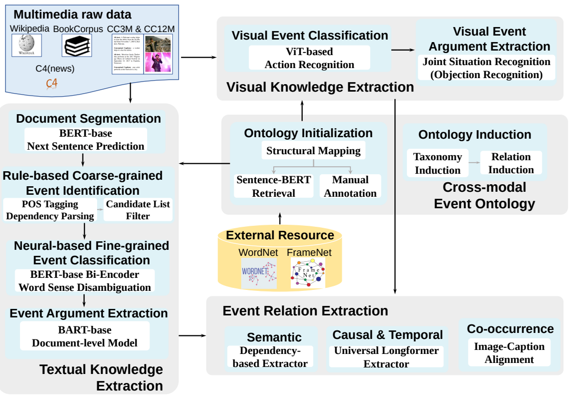


Figure 5: The architecture of Extraction System. There are four main components: Cross-modal Event Ontology, Textual Knowledge Extraction, Visual Knowledge Extraction, and Event Relation Extraction. We use the same ontology introduced in Section 3 for both MMEKG and this extraction system. Another three components are introduced in Section 4 respectively.

Relation Induction aims to discover commonsense relations between concept events, based on the relations between instance events. Similar to taxonomy induction, we calculate a salience score $s_r(o_h, o_t)$ for each pair of concept events (o_h, o_t) on relation r. The score considers (1) the confidence score of relation r between the children instance events. (2) the commonality of o_t w.r.t. r. We add (o_h, r, o_t) with a salience score exceeding a threshold to MMEKG. For example in Figure 3, since the salience score of the triple (talk.v.01, *co-occur*, sit.v.01) exceeds the threshold, we expand such relation from instance-level to concept-level.

4 Knowledge Extraction System

This section briefly introduces our knowledge extraction system collecting large-scale instance events and relations for MMEKG, which is shown in Figure 5. We follow the overall framework of previous knowledge extraction systems like GAIA (Li et al., 2020b) and RESIN (Wen et al., 2021), but extends and optimizes event-related components to enable it extracting billion-scale, high-quality events efficiently. With more advanced models, tuning strategy and component architectures, our system achieves comparable if not better performance on each component using a common benchmark. We also substitute all Cross-encoder in the system to Bi-encoder if possible and conduct a joint model of multi-task training during event relation extraction for efficiency.

4.1 Textual Knowledge Extraction

This component extracts nodes of the event graph from unstructured texts via event detection and argument extraction. (1) We **pre-process** the corpus as follows. First, we identify document boundaries using BERT-base Next Sentence Prediction (NSP) model and heuristic rules (5-10 sentences per document). Then, we obtain POS-tag and dependency tree via Stanza (Qi et al., 2020). Verbs and adjectives are regarded as candidate words triggering events. (2) Thanks to the synsets in our ontology, we convert **Event Detection** as an unsupervised word sense disambiguation (WSD) task to avoid costly training data. We apply a Bi-encoder model (Blevins and Zettlemoyer, 2020) to predict the most possible synset for candidate trigger words. Each synset refers to a concept event. We thus can link the texts with MMEKG. (3) We propose an efficient and effective method named PAIE (Ma et al., 2022) for **Event Argument Extraction**. The basic idea is to extend QA-based models (Du and Cardie, 2020) to predict all roles for a target event simultaneously. We propose to prompt PLMs for extraction tasks and design a role interaction prompt template for each concept event. All role embeddings serve as query vectors to identify argument spans as the answer. We train the model on annotations provided by FrameNet.

4.2 Visual Knowledge Extraction

For visual knowledge extraction, we design a two-stage extraction network. Both models are trained using the largest visual situation recognition dataset (Yatskar et al., 2016; Pratt et al., 2020). (1) For event detection, we leverage pre-trained ViT (Dosovitskiy et al., 2021) to obtain patched image features. Then, another layer of transformer is finetuned to classify images into our visual concept events. (2) Following Pratt et al. (2020), we use pre-trained ResNet-50 (He et al., 2016) as the backbone of Faster R-CNN (Ren et al., 2015), and conditional LSTM decoder to aggregate role information to extract arguments from images.

4.3 Event Relation Extraction

This component aims to extract temporal, causal, co-occur, and semantic relations between instance events. Co-occur includes text/image alignments.

Temporal and Causal Relation. For temporal and causal relations, we propose a novel method that builds a document-level graph to infer the relations

Component	Sub-task	Benchmark	Metric	Our score	SOTA
Text Event Extraction	WSD	SemEval-2007	Accuracy	74.5	77.4 (Barba et al., 2021)
	EAE	ACE-05	F1	67.0	65.4 (Du and Cardie, 2020)
Visual Event Extraction	VerbD	imSitu	Accuracy	46.8	43.2 (Suhail and Sigal, 2019)
	EAE	imSitu	Accuracy	23.8	19.5 (Suhail and Sigal, 2019)
Event Relation Extraction	ECI	Causal-TimeBank	F1	61.7	53.2 (Zuo et al., 2021)

Table 1: Performance of each component. Abbreviation in column **Sub-task**: EAE: Event Argument Extraction. VerbD: Verb Detection. ECI: Event Causality Identification.

	#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

Table 2: Statistics of MMEKG and existing event KG.

Table 3: Taxonomy induction.

#Sample	Positive	Negative
1000	80.1%	19.9%

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

Table 4: Instance-level evaluation.

among events globally. Our method could conduct across-sentence reasoning without clear temporal/causal indicators and complicated heuristic rules. This enables us to identify all temporal and causal relations of a document simultaneously and efficiently. We jointly predict temporal and causal relations as multi-label multi-task classification and train the model based on Causal-TimeBank (Mirza, 2014). There are six relation types in total: *Before*, *After*, *During*, *Includes*, *Included*, and *Causal*.

Co-occurrence Relation. For textual co-occurrence, we identify it via dependency parsing if the trigger words have a *conj* relation. For cross-modal co-occurrence, we extract events from paired image-caption respectively and assume they co-occur. We also observe semantic shifts between different modalities. As shown in Figure 1, the textual *dressing* event may be a sub-event of the visual *sleeping* event. We will investigate it soon.

Semantic Relation. We claim that when an argument of event A is a gerund phrase B, B could also be viewed as a sub-event of A triggered by the gerund functioning as its semantic component. For example, we extract two events from sentence *Eating too much fried chicken cause overweight*: cause overweight (event A) and eat too much chicken (event B). Since A is also an argument of role *influencing_entity* for B, event eat too much chicken and cause overweight are connected with relation *influencing_entity*. Based on such assumption, we expand the relation types by exploiting the *frame elements* in FrameNet. We capture all event pairs in sentences satisfying (1) the trigger words are connected by *acl* or *acl:relcl* in dependency parsing, or (2) the trigger of one event is extracted as an argument of another event. Then we identify these two events having a relation labeled by the argument role.

5 Evaluation

5.1 MMEKG Statistics

Table 2 presents the statistics of MMEKG and other Event KGs. We build a full version, MMEKG-full, and MMEKG-core which filters out infrequent events ($<$ 3 times), leading to a denser and more accurate version. MMEKG involves not only a much larger ontology but also more instance events.

5.2 Extraction System Performance

Table 1 shows the results of our components trained on publicly available datasets, since there is no unified benchmark to evaluate the entire extraction process. We can see that all of our knowledge extraction components, except WSD, achieve better performance. Our WSD model performs comparably and efficiently for massive event detection.

5.3 Instance-level Evaluation

Considering the different data distribution between training data and extracted corpus, we manually evaluate the instance-level quality of MMEKG. We randomly select 1,000 instance events in texts and

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%

Table 5: Relation induction.

500 from images. Along with original contexts, we invite six colleagues to label whether the extracted event represents the semantic meaning of the original source or not. For instance event relations, we consider: (1) causal/temporal relations from texts and (2) cross-modal co-occurrence from image-caption pairs. We sample 200 textual relations and 300 cross-modality relations. Along with the contexts, we provide these extracted relations to the same six colleagues and ask them whether the relation extracted matches the original resource. Results in Table 4 demonstrate little performance degradation in precision[5] and an acceptable quality of our proposed MMEKG, considering the complexity of the entire pipeline.

5.4 Ontology-level Evaluation

Large-scale ontology is critical for knowledge reasoning. We further evaluate the quality of inferred taxonomy and relations. The difference from the instance-level evaluation is that no context is provided for reference in ontology evaluation. We construct pairs with one positive and one negative sample for comparison convenience, as illustrated in Figure 6, and ask the same six colleagues which sample agrees with our commonsense more. The results are shown in Tables 3 and 5. Both negatives are around 20%. In particular, for relation induction, some similar pairs are hard to tell which one is better. We attribute this to the low recall and random negative sampling, which may bring in false negatives. This also provides insights for future improvements.

6 Related Work

Event Knowledge Graph Existing event knowledge graphs (Speer et al., 2016; Sap et al., 2019; Zhang et al., 2020) usually face a dilemma about quality and quantity. ATOMIC (Sap et al., 2019) annotates manually and constructs high-quality knowledge bases, while ASER (Zhang et al., 2020) leverages defined patterns and automatic pipeline to build a large-scale graph. Compared with ASER, we not only develop a larger KG by larger corpus and advanced extraction system but also derive complicated ontology and incorporate information across modalities to control the quality of KG.

```
Taxonomy Induction:
 positive: eat.v.01__Ingestibles:apple
 negative: eat.v.01__Ingestibles:table

Relation Induction:
 positive: (
   destroy.v.02__Patient:building, include,
     kill.v.01___Victim:people )
 negative: (
   destroy.v.02__Patient:building, include,
     invalidate.v.01__Phenomenon:subpoena )
```

Figure 6: Examples of pair constructed for taxonomy (top) and relation induction (bottom). Each pair includes one positive and one negative sample. Positive ones are sampled from induced concept events or relations. Negative ones are generated by substituting the arguments (taxonomy) or tail events (relation) in positive samples.

Knowledge Extraction System Previous multi-modal knowledge extraction systems, such as GAIA (Li et al., 2020b) and RESIN (Wen et al., 2021), jointly extract information of a small domain from relatively small-scale resource. Our system inherits their overall framework but is applied for extracting billion-scale and universal events. Therefore we optimize event-related modules targetedly for both efficiency and effectiveness.

Cross-media Event Argument Alignment Some previous works (Li et al., 2020a; Fung et al., 2021) also bridge texts and images through fine-grained alignments of event arguments for various tasks, such as multi-modal event extraction and fake news detection. Instead, we fuse knowledge from different modalities to construct such a large-scale KG.

7 Conclusion

We present the first Multi-modal Event KG (MMEKG) with a large-scale event ontology. It not only bridges and complements different modalities of knowledge via more expressive events but also benefits comprehensive reasoning with rich cross-modal contexts. Additionally, we provide a demo system that can seamlessly parse and link any texts/images via our knowledge extraction system.

[5] We do not report recall here because (1) there is no annotated ground truth of instance events in extracted corpus. (2) precision is much more important since we extract events from large corpus to construct KG.

Acknowledgments

This study is supported under the RIE2020 Industry Alignment Fund – Industry Collaboration Projects (IAF-ICP) Funding Initiative, as well as cash and in-kind contribution from the industry partner(s). We also thank the KULeuven C1 project Macchina for support.

References

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The Berkeley FrameNet project. In *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 1*, pages 86–90, Montreal, Quebec, Canada. Association for Computational Linguistics.

Edoardo Barba, Luigi Procopio, and Roberto Navigli. 2021. ConSeC: Word sense disambiguation as continuous sense comprehension. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1492–1503, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Terra Blevins and Luke Zettlemoyer. 2020. Moving down the long tail of word sense disambiguation with gloss informed bi-encoders. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1006–1017, Online. Association for Computational Linguistics.

Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. 2021. Conceptual 12M: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *CVPR*.

Xin Dong, Evgeniy Gabrilovich, Geremy Heitz, Wilko Horn, Ni Lao, Kevin Murphy, Thomas Strohmann, Shaohua Sun, and Wei Zhang. 2014. Knowledge vault: A web-scale approach to probabilistic knowledge fusion. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 601–610.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*.

Xinya Du and Claire Cardie. 2020. Event extraction by answering (almost) natural questions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 671–683, Online. Association for Computational Linguistics.

Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. Bradford Books.

Yi Fung, Christopher Thomas, Revanth Gangi Reddy, Sandeep Polisetty, Heng Ji, Shih-Fu Chang, Kathleen McKeown, Mohit Bansal, and Avi Sil. 2021. InfoSurgeon: Cross-media fine-grained information consistency checking for fake news detection. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1683–1698, Online. Association for Computational Linguistics.

Simon Gottschalk and Elena Demidova. 2019. Eventkg–the hub of event knowledge on the web–and biographical timeline generation. *Semantic Web*, 10(6):1039–1070.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Jena D. Hwang, Chandra Bhagavatula, Ronan Le Bras, Jeff Da, Keisuke Sakaguchi, Antoine Bosselut, and Yejin Choi. 2021. Comet-atomic 2020: On symbolic and neural commonsense knowledge graphs. In *AAAI*.

Tuan Lai, Heng Ji, ChengXiang Zhai, and Quan Hung Tran. 2021. Joint biomedical entity and relation extraction with knowledge-enhanced collective inference. In *ACL*.

Svetlozara Leseva and Ivelina Stoyanova. 2019. Structural approach to enhancing WordNet with conceptual frame semantics. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 629–637, Varna, Bulgaria. INCOMA Ltd.

Lily Li, Or Levi, Pedram Hosseini, and David Broniatowski. 2020a. A multi-modal method for satire detection using textual and visual cues. In *Proceedings of the 3rd NLP4IF Workshop on NLP for Internet Freedom: Censorship, Disinformation, and Propaganda*, pages 33–38, Barcelona, Spain (Online). International Committee on Computational Linguistics (ICCL).

Manling Li, Alireza Zareian, Ying Lin, Xiaoman Pan, Spencer Whitehead, Brian Chen, Bo Wu, Heng Ji, Shih-Fu Chang, Clare Voss, Daniel Napierski, and Marjorie Freedman. 2020b. GAIA: A fine-grained multimedia knowledge extraction system. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 77–86, Online. Association for Computational Linguistics.

Kang Liu, Yubo Chen, Jian Liu, Xinyu Zuo, and Jun Zhao. 2020. Extracting event and their relations from texts: A survey on recent research progress and challenges. *AI Open*, 1:22–39.

Yubo Ma, Zehao Wang, Yixin Cao, Mukai Li, Meiqi Chen, Kun Wang, and Jing Shao. 2022. Prompt for extraction? paie: Prompting argument interaction for event argument extraction. *arXiv preprint arXiv:2202.12109*.

Brian McBride. 2004. The resource description framework (rdf) and its vocabulary description language rdfs. In *Handbook on ontologies*, pages 51–65. Springer.

Paramita Mirza. 2014. Extracting temporal and causal relations between events. In *Proceedings of the ACL 2014 Student Research Workshop*, pages 10–17, Baltimore, Maryland, USA. Association for Computational Linguistics.

Sarah Pratt, Mark Yatskar, Luca Weihs, Ali Farhadi, and Aniruddha Kembhavi. 2020. Grounded situation recognition. *ArXiv*, abs/2003.12058.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108, Online. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28:91–99.

Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A Smith, and Yejin Choi. 2019. Atomic: An atlas of machine commonsense for if-then reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 3027–3035.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565.

Robert Speer, Joshua Chin, and Catherine Havasi. 2016. Conceptnet 5.5: An open multilingual graph of general knowledge. In *AAAI Conference on Artificial Intelligence*.

Mohammed Suhail and Leonid Sigal. 2019. Mixture-kernel graph attention network for situation recognition. In *The IEEE International Conference on Computer Vision (ICCV)*, pages 10363–10372.

Willem Robert Van Hage, Véronique Malaisé, Roxane Segers, Laura Hollink, and Guus Schreiber. 2011. Design and use of the simple event model (sem). *Journal of Web Semantics*, 9(2):128–136.

Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, 57(10):78–85.

Haoyang Wen, Ying Lin, Tuan Lai, Xiaoman Pan, Sha Li, Xudong Lin, Ben Zhou, Manling Li, Haoyu Wang, Hongming Zhang, Xiaodong Yu, Alexander Dong, Zhenhailong Wang, Yi Fung, Piyush Mishra, Qing Lyu, Dídac Surís, Brian Chen, Susan Windisch Brown, Martha Palmer, Chris Callison-Burch, Carl Vondrick, Jiawei Han, Dan Roth, Shih-Fu Chang, and Heng Ji. 2021. RESIN: A dockerized schema-guided cross-document cross-lingual cross-media information extraction and event tracking system. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Demonstrations*, pages 133–143, Online. Association for Computational Linguistics.

Mark Yatskar, Luke Zettlemoyer, and Ali Farhadi. 2016. Situation recognition: Visual semantic role labeling for image understanding. In *Conference on Computer Vision and Pattern Recognition*.

Hongming Zhang, Xin Liu, Haojie Pan, Yangqiu Song, and Cane Wing-Ki Leung. 2020. Aser: A large-scale eventuality knowledge graph. In *Proceedings of The Web Conference 2020*, pages 201–211.

Xinyu Zuo, Pengfei Cao, Yubo Chen, Kang Liu, Jun Zhao, Weihua Peng, and Yuguang Chen. 2021. Improving event causality identification via self-supervised representation learning on external causal statement. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2162–2172, Online. Association for Computational Linguistics.