mathematics

MDPI

Article

A Survey on Multimodal Knowledge Graphs: Construction, Completion and Applications

Yong Chen [1], Xinkai Ge [2,*], Shengli Yang [3], Linmei Hu [4,*], Jie Li [2] and Jinwen Zhang [5]

[1] School of Computer Science and Technology, University of Science and Technology of China, Hefei 230052, China
[2] The Institute of Computer Science, Beijing University of Posts and Telecommunications, Beijing 100876, China
[3] National Security School, China People's Liberation Army National Defence University, Beijing 100091, China
[4] School of Computer Science, Beijing Institute of Technology, Beijing 100811, China
[5] North Automatic Control Technology Institute, Taiyuan 030006, China
* Correspondence: gexinkai2017@163.com (X.G.); hulinmei@bit.edu.cn (L.H.)

Abstract: As an essential part of artificial intelligence, a knowledge graph describes the real-world entities, concepts and their various semantic relationships in a structured way and has been gradually popularized in a variety practical scenarios. The majority of existing knowledge graphs mainly concentrate on organizing and managing textual knowledge in a structured representation, while paying little attention to the multimodal resources (e.g., pictures and videos), which can serve as the foundation for the machine perception of a real-world data scenario. To this end, in this survey, we comprehensively review the related advances of multimodal knowledge graphs, covering multimodal knowledge graph construction, completion and typical applications. For construction, we outline the methods of named entity recognition, relation extraction and event extraction. For completion, we discuss the multimodal knowledge graph representation learning and entity linking. Finally, the mainstream applications of multimodal knowledge graphs in miscellaneous domains are summarized.

Keywords: multimodal knowledge graph; knowledge graph construction; knowledge graph completion; multimodal knowledge graph application

MSC: 68T30

Citation: Chen, Y.; Ge, X.; Yang, S.; Hu, L.; Li, J.; Zhang, J. A Survey on Multimodal Knowledge Graphs: Construction, Completion and Applications. *Mathematics* **2023**, *11*, 1815. https://doi.org/10.3390/math11081815

Academic Editors: Tianxing Wu, Yuxiang Wang and Ningyu Zhang

Received: 3 March 2023
Revised: 23 March 2023
Accepted: 30 March 2023
Published: 11 April 2023

1. Introduction

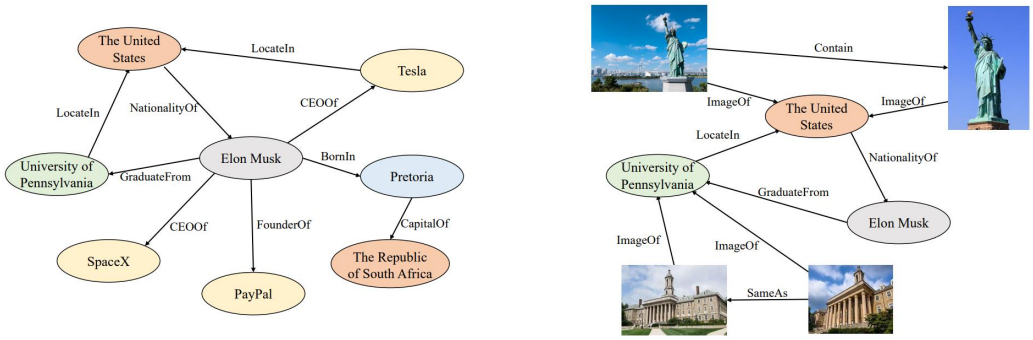
As a successful application of knowledge engineering in big data, knowledge graph describes the concepts, entities and their relationships in a structured form. A knowledge graph can be viewed as a structured representation of facts that can be expressed in a factual triple in the form of head, predicate and tail) under the RDF [1], with head and tail being entities and predicate being the relation type. In addition to triples, a knowledge graph can also be represented as a multirelational graph, where nodes represent entities and directed edges represent relationships.

Ranging from general to domain-specific purposes, knowledge graphs contribute to organizing, managing and understanding the massive information on the Internet and facilitate the development of intelligent services, such as recommender systems [2], dialogue systems [3], semantic search [4] and other miscellaneous systems. A general knowledge graph can be visually regarded as a "structured encyclopedia knowledge base" for general fields, which contains a large number of common sense knowledge in the real world [5]. They not only contain a large amount of semistructured and unstructured data but also have high domain coverage. Representative works of general knowledge graphs over the last years include Freebase [6], DBpedia [7] and Wikidata [8]. On the

other hand, domain-specific knowledge graphs, also called industry knowledge graphs or vertical knowledge graphs, are usually applied to a specific field and can be regarded as an "industry knowledge based on semantic technology" whose construction relies on the data of a specific industry. Typical domain-specific knowledge graphs include IMDB (http://www.imdb.com, 1 March 2022), MusicBrainz (http://musicbrainz.org/, 1 March 2022) and UMLS [9].

Nevertheless, the above-mentioned knowledge graphs focus on the textual facts with few multimodal sources. Actually, in addition to text and structured data, visual and auditory data, such as pictures, videos and audio, can also be the data sources. These different data sources complement and strengthen each other when describing the same object, thus improving the performance of knowledge graph tasks over unimodal models and facilitating the machine's perception of the real data scenarios [10]. In light of this, we view a multimodal knowledge graph as a graph of data intended to accumulate and convey knowledge from multimodal views such as textual, visual and auditory views. A well-constructed multimodal knowledge graph can provide a wider data scope and research base for researchers from natural language processing and computer vision, and further promote cross-domain fusion research.

Recent advances in multimodal knowledge graphs focus on construction [11] and completion [12,13]. Due to the proliferation of visual resources on the Web, current multimodal knowledge graph construction mostly focuses on textual and visual resources. Examples of traditional knowledge graphs and multimodal knowledge graphs are illustrated in Figure 1. Typical multimodal knowledge graphs include IMGpedia [14], MMKG [15] and Richpedia [16]. Related research has been successfully applied in miscellaneous tasks, such as multimodal entity linking [17], recommender system [18] and E-commerce [19].



(a) Knowledge Graph (b) Multimodal Knowledge Graph

Figure 1. An example of a knowledge graph and a multimodal knowledge graph.

In this paper, we provide a comprehensive survey of multimodal knowledge graphs including construction, completion and typical applications in different domains. In particular, we focus on multimodal knowledge graphs based on textual and visual data resources. The contributions of this survey are twofold. First, we comprehensively summarize the development and typical examples of multimodal knowledge graphs. In addition, the multimodal knowledge graph construction as well as completion techniques are systematically introduced and organized. Second, we provide a concrete taxonomic schema to organize the multimodal knowledge graph construction and completion technologies. Specifically, we present the graph construction technologies including named entity recognition, relation extraction and event extraction. Analogously, we survey the knowledge completion technologies which include entity linking and knowledge representation learning. The taxonomy of this survey is described in Figure 2.

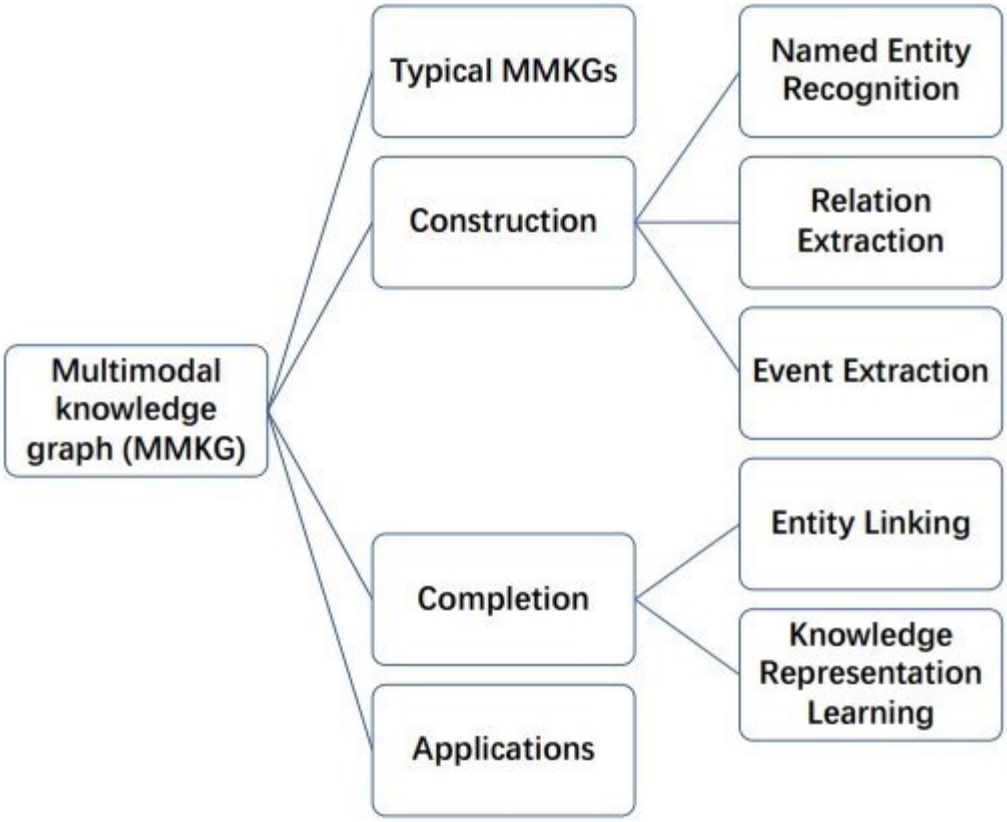


Figure 2. The taxonomy of the survey on MMKG.

2. Open Multimodal Knowledge Graphs

A traditional knowledge graph $\mathcal{G}$ can be defined as $\mathcal{G} = \{\mathcal{E}, \mathcal{R}, \mathcal{F}\}$ where $\mathcal{E}$, $\mathcal{R}$ and $\mathcal{F}$ represent the collection of entities, relations and facts, respectively. Facts are composed of triples, (*h, r, t*) where *h*, *r* and *t* represent head entity, relation and tail entity, respectively. According to the definition of traditional knowledge graph [20], a multimodal knowledge graph $\mathcal{G}$ can be seen as a knowledge graph whose entities $\mathcal{E}$ are associated with data in modalities other than text (e.g., images). Existing multimodal knowledge graphs mainly adopt two different ways for representing visual information. One way is to represent multimodal data as particular attribute values of entities, while the other way takes multimodal data as entities, which are associated with the corresponding concepts through specific types of relations. [20] In this section, we enumerate and introduce some existing typical multimodal knowledge graphs.

2.1. MMKG

Knowledge Graphs have been used as external sources of knowledge for multifarious tasks. As a result, much research has focused heavily on the problem of knowledge graph completion. Considering that it is one-sided to evaluate completion approaches on only one knowledge graph [21], Liu et al. [15] proposed MMKG to address the issues. Contrary to previous knowledge graphs, MMKG contains both numerical features and images for all entities as well as entity alignments between pairs of knowledge graphs, which is specially designed for tackling link prediction and entity matching problems. The aim of the former is to infer missing links within the KG while that of the latter is to find pairs of records which refer to the same entity.

MMKG created two knowledge graphs DBpedia15K and YAGO15K, which are the counterparts of DBpedia and YAGO, respectively. In terms of specific construction process, MMKG employs Freebase15K [22] as the blue print for the construction of multimodal knowledge graphs, based on which the construction of MMKG can be roughly summarized

as the following steps: Firstly, by entity alignment with DBpedia [7] and YAGO [23] through the sameAs links, respectively, MMKG incorporates more entities for downstream tasks. Numeric literals, which are linked by numerical relations such as the relation */location/geocode/latitude*, with which entities are then integrated, after which MMKG collects images through a web crawler aiming to parse query results for the image search engines, such as Google Images, Bing Images and Yahoo Image Search, and takes them as attribute nodes of each entity. To minimize the noise caused by polysemous entity labels, the authors adopted Wikipedia URIs as query strings which are processed and used as search queries for disambiguation purposes. For example, for the entity "Paris", we can obtain URIs such as *Paris (ile-de-France, France)* and *Paris (City of New Orleans, Louisiana)*. With the above procedures, MMKG stores 55.8 images per entity on average. Concrete construction details can be found in MMKG [15].

2.2. IMGpedia

To address the problem that the existing datasets describing multimedia are focused on capturing the metadata of the multimedia files rather than the multimedia content itself, Ferrada et al. [14] proposed IMGpedia, which incorporates visual descriptors and visual similarity relations for the images of Wikimedia Commons dataset and links them with both DBpedia Commons dataset and DBpedia dataset for a variety of applications, such as visual similarity calculation and visual–semantic queries over the images.

In terms of the construction of IMGpedia, concretely, IMGpedia first gathers about 14.7 million images from Wikimedia Commons and then proceeds to compute different visual descriptors to capture different elements of the content of the images, such as Gray Histogram Descriptor, Histogram of Oriented Gradients Descriptor and Color Layout Descriptor. Aided by visual descriptors, IMGpedia can obtain the visual similarity between pairs of images. Thereafter, IMGpedia represents this information as an RDF through a custom lightweight ontology. Through the above steps, a dataset of IMGpedia containing information about 14.7 million images of Wikimedia Commons, the description of their content as well as links to their most similar images and to the DBpedia resources that form part of their context is established, which supports queries based on SPARQL query and is available for many potential use-cases.

2.3. Richpedia

In order to solve the lack of complete multimodal graphs in the academic community, which hinders future research on multimodal fusion, Wang et al. [16] infused visual–relational resources into general knowledge graphs and established a multimodal knowledge graph, Richpedia.

Overall, Richpedia is a finite set of Richpedia triples, t = <*subject*, *predicate*, *object*>, where t is a member of set $(\mathcal{E} \cup \mathcal{B}) \times \mathcal{R} \times (\mathcal{E} \cup \mathcal{L} \cup \mathcal{B})$. Richpedia defines relations between entities as $\mathcal{R}$ and the entity set as $\mathcal{E} = \mathcal{E}_{\mathcal{KG}} \cup \mathcal{E}_{\mathcal{IM}}$, where $\mathcal{E}_{\mathcal{KG}}$ is general KG entities and $\mathcal{E}_{\mathcal{IM}}$ is image entities. $\mathcal{L}$ and $\mathcal{B}$ represent the set of literals and blank nodes, respectively.

Richpedia was constructed in three steps: data collection, image processing and relation discovery. In the first stage, general KG entities and image entities were collected respectively. Richpedia mainly collected 30,638 KG entities regarding cities, sights and celebrities from Wikidata, and 2,883,162 images entities from Wikipedia and web sources aided by the web crawler. During the second step, since the images collected from web search engines may be duplicated or irrelevant to KG entities, the authors chose K-means algorithm on visual features (such as gradient histogram descriptor, color layout descriptor, etc.) extracted by the VGG16 to filter out the noise image entities. After the denoising process, visual descriptors are introduced to calculate the similarity between images by integrating the distance between different descriptors. Last of all, the relation discovery between image entities for Richpedia was conducted. Due to the existence of potential relations between scattered image entities (e.g., "imageof", "sameAs"), the authors propose three effective rules to extract and infer these semantic relationships from

unstructured information about image entities, namely, relevant hyperlinks and text in Wikipedia. Subsequently, the authors created a custom lightweight Richpedia ontology to represent the data in RDF format.

Richpedia (http://richpedia.cn, 1March 2022) provides a facet query endpoint to allow researchers to retrieve and leverage data distributed over general KGs and image resources to answer richer visual queries and make multirelational link predictions.

2.4. ImageGraph

ImageGraph [24], based on FB15K, is a multirelational graph within which images are introduced and associated with the corresponding entities as attributes. In total, it contains 1330 relation types, 14,870 entities and 829,931 images crawled from the web.

The construction of ImageGraph can be summarized into the following steps: First, as FB15K does not contain any visual data, a web crawler is applied to obtain images from Google Images, Bing Images and Yahoo Image Search. To minimize polysemous entity labels brought by noise, Wikipedia URIs are used for disambiguation. For example, there are more than 100 Freebase entities with the label "Springfield" to distinguish these entities, and URIs like $Springfield\_(Massachusetts, United\_States)$ and $Springfield\_(MA)$ are used (these two URIs specify the entity representing Springfield, Massachusetts). Then, the corrupted, low-quality and duplicate images are removed from the crawled images, and only the top 25 images are chosen as the associated images for each entity, with the images scaled to a maximum height or width of 500 pixels while keeping their aspect ratio. In the end, triples containing a head or tail entity that could not be associated with an image are filtered out, and the ImageGraph dataset is obtained.

Numerous applications can benefit from visual–relational KGs like ImageGraph. They promote numerous novel query types through introducing images to be arguments of queries, bringing more efficient and accurate query answering.

2.5. VisualSem

Alberts et al. [25] released VisualSem, which is a high-quality knowledge graph containing nodes with multilingual glosses and multiple illustrative images, where nodes represent concepts and named entities with well-curated related images as well as glosses in multiple languages as attributes. Specifically, VisualSem consists of 89,896 nodes with 1,342,764 glosses and 938,100 images.

To gather information from various sources, BabelNet API (https://babelnet.org/guide, 1 March 2022) [26], a big multilingual and multimodal resource that compiles data from a variety of sources, is applied for constructing the knowledge graph during the construction process. Specifically, the construction process starts from choosing a set of initial nodes which can guarantee the high-quality images associated with them. As ImageNet classes satisfy the above conditions, BabelNet API is used to obtain synsets corresponding to the 1000 ImageNet classes used in the ILSVRC image classification competition [27], which forms the initial nodes. These initial nodes are referred as the initial node pool. With initial nodes set, an iterative process containing four consecutive steps are performed for construction. Firstly, neighbor nodes are retrieved for each node in the node pool through BabelNet API. Namely, all first-degree nodes are picked up using BabelNet API, with duplicate ones being removed. Secondly, the linked images ware validated, and those do not meet the predefined quality standards are removed. Thirdly, the CLIP model [28] is applied to further filter images that do not match the corresponding nodes' glosses based on a predefined threshold. Lastly, the top-k nodes among remaining nodes are accepted after sorting. The above four steps are processed iteratively until the number of nodes reaches 90,000.

VisualSem is designed to be used in vision and language research and can be easily integrated into neural model pipelines, which has the potential to facilitate various sorts of natural language understanding (NLU) and natural language generation (NLG) tasks in data augmentation or data grounding settings.

3. Multimodal Knowledge Graph Construction

In this section, we provide a comprehensive review of the research of multimodal knowledge graph construction, namely knowledge acquisition, which aims to discover and recognize entities and relations from text sources as well as sources from other modalities. In this section, we divide the knowledge acquisition techniques into three categories: entity recognition, relation extraction and event extraction.

3.1. Named Entity Recognition

Named entity recognition (NER) is designed to recognize and classify named entities within natural texts and visual objects within images or videos into predefined categories such as person, location, organization, etc. [29]. Besides acting as an essential role in information extraction, NER is also widely used in various downstream applications, such as information retrieval [30,31], knowledge base construction [32], question answering [33] and machine translation [34].

This section briefly summarizes the text-based named entity recognition and introduces the multimodal named entity recognition in detail.

3.1.1. Text-based Named Entity Recognition

Previous methods of NER have usually been based on hand-crafted rules, such as entity dictionaries and word forms. With the rise of deep learning, deep-learning methods, which are conducive to the automatic discovery of hidden features, have become the mainstream.

Traditional methods for NER can be divided into three streams: rule-based, unsupervised learning and feature-based supervised learning methods. (1) Rule-based methods mainly rely on hand-crafted rules, which can work well when the lexicon is exhaustive. Typical rule-based NER systems include LaSIE-II [35], SAR [36], FASTUS [37] and LTG [38] systems. Nevertheless, these systems have poor transferability due to the domain-specific rules and incomplete dictionaries. (2) Feature-based supervised learning methods are more widely used compared with the above methods. Given data samples to be annotated, these methods cast NER to a multiclass classification or sequence labeling task. Common methods include hidden Markov models (HMM) [39,40], decision trees [41,42] and conditional random fields [43–45].

Compared with traditional methods, the deep-learning-based methods are useful in discovering hidden features automatically and have achieved superior results. Deep-learning-based methods can be further divided into three categories: (1) CNN-based models are widely used for NER tasks. Wu et al. [46] adopted a sentence-level log-likelihood approach [47], consisting of a convolutional layer, a nonlinear layer and several linear layers for NER. Strubell et al. [48] adopted an iterated dilated CNN architecture to incorporate global information from a whole sentence. (2) RNN-based models have achieved remarkable achievements in NER. Huang et al. [49] were first to utilize bidirectional LSTM CRF architecture on sequence tagging tasks. Zhang et al. [50] leveraged a lattice LSTM structure to automatically choose the most relevant characters and words for better NER results. (3) The pretrained language model is becoming a new paradigm in NER with the superior performance of BERT [51]. Figure 3 compares the differences between BiLSTM and BERT in named entity recognition. Zhang et al. [52] adopted the pretrained BERT model to obtain expressive sentence features for resolving the problem of limited labeled resources and domain shift in NER. Liu et al. [53] applied BERT to extract underlying features of texts for conducting NER in the biomedical field. Fang et al. [54] exploited BERT-based character vectors and embedded them into a deep learning model for performing Chinese NER.

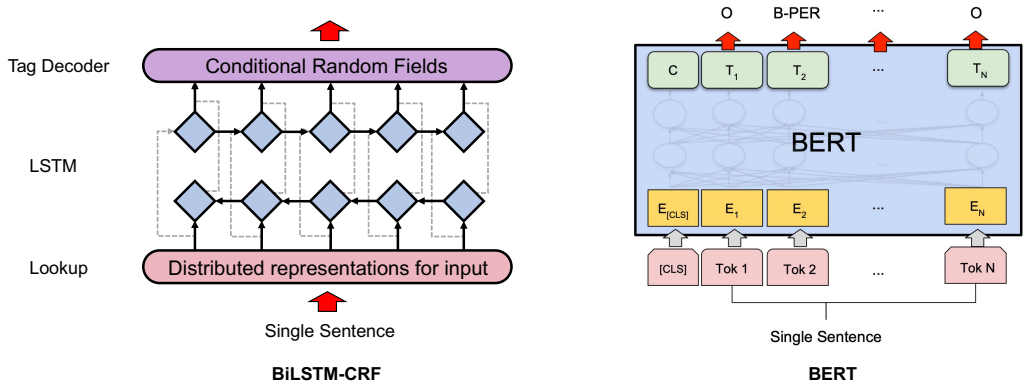


Figure 3. Comparison of BiLSTM and BERT in text-based named entity recognition.

3.1.2. **Multimodal Named Entity Recognition**

Most unstructured texts often do not provide a sufficient textual context to resolve polysemous entities and may contain a massive number of unknown tokens. As shown in Figure 4, with visual information, we can know Rocky is a dog instead of a person named Rocky. To address these challenges, Moon et al. [55] proposed the multimodal named entity recognition task, which aims at exploiting practical visual information to improve the performance of NER. We can further classify the multimodal named entity recognition into direct concatenation-based methods, gated fusion-based methods, multimodal alignment-based methods, pretrained multitask-based methods and graph-based methods according to the way in which information from different modalities is fused.

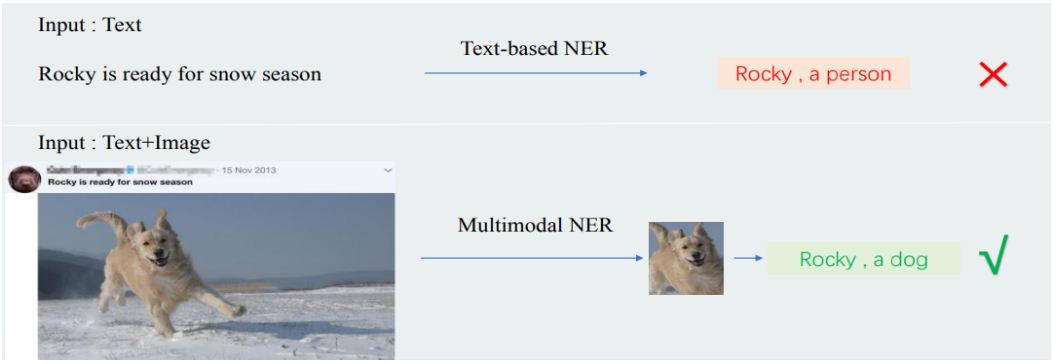


Figure 4. Illustration of multimodal named entity recognition.

Direct Concatenation-based Methods. Some methods apply simple concatenation when fusing multimodal information. Moon et al. [55] proposed a BiLSTM-CNN hybrid model to extract a relevant visual context to augment textual information for the recognition of a named entity in the text input. Zhang et al. [56] proposed an adaptive coattention network to integrate the multimodal features extracted with BiLSTM and VGG to recognize named entities. Specially, in addition to the visual clues, Shahzad et al. [57] also considered text information under the level of word, character and sentence for NER in short text.

Gated Fusion-based Methods. Some methods introduce the gated mechanism to filter the key information. Lu et al. [58] adopted an attention-based model to extract those visual features most related to the text and employed a visual gate to control the combination of visual features and text representation generated by BiLSTM for multimodal NER. Analogously, Arshad et al. [59] extended the self-attention mechanism to capture relationships between two words and image regions, and further introduced a gated fusion module to dynamically select information from multimodal features.

Multimodal Alignment-based Methods. Some methods focus on the alignment between information from different modalities. Wu et al. [60] utilized the pretrained Mask-RCNN [61] to extract the visual objects and introduced a dense coattention network to

model the correlations between visual objects and textual entities as well as the internal connections of objects or entities. Similarly, Zheng et al. [62] introduced gated bilinear attention to capture the mapping relations between visual objects detected by Mask-RCNN [61] and textual entities. In order to map two different representations into a shared representation, they adopted a strategy of adversarial training for a better fusion of the two modalities to improve the performance of multimodal NER. Asgari-Chenaghlu [63] adopted BERT to integrate the textual features and visual features extracted by the InceptionV3 model [64] to recognize named entities. Based on [65], Sun et al. [66] upgraded the original BiLSTM to BERT and introduced a method of text-image relation propagation for multimodal NER. In order to augment the interaction between modalities and alleviate the visual bias caused by the dataset, Yu et al. [67] proposed a unified multimodal architecture based on transformer to capture the implicit alignments between words and images. Furthermore, they leveraged a text-based entity span detection module to largely eliminate the bias of the visual context.

Pretrained Multitask-based Methods. Some methods apply multiple tasks for pretraining their model so as to fuse multimodal information. Sun et al. [65] designed a text–image relation classification task and a next-word prediction task for pretraining a multimodal language model. Using a semisupervised paradigm and a multitask framework, their model can resolve the problem of inappropriate visual clues fused in the multimodal model, which causes a negative impact on multimodal NER.

Graph-based Methods. Zhang et al. [68] represented the sentence and image as a multimodal graph, with each node indicating the textual word or visual object. By constructing a unified multimodal graph fusion approach, their model can capture various semantic relationships between words and objects to perform entity labeling.

3.2. Relation Extraction

Relation extraction (RE) is one of the crucial techniques involved in information extraction. It refers to extracting new relation facts between entities from plain text and adding them into knowledge graphs. Most existing research on relation extraction exploits neural-network-based approaches. Nevertheless, these methods are mainly text-based and suffer poor performance when texts lack contexts. In light of this, Zheng et al. [69] found that image-related information can supplement the missing contexts in social media texts. Thus, they proposed multimodal relation extraction, which refers to classifying textual relations between two entities with the help of visual contents.

This section provides a brief overview of text-based relation extraction and a detailed summary of the multimodal relation extraction approaches developed thus far.

3.2.1. Text-Based Relation Extraction

Supervised methods based on CNN-based models and RNN-based models were dominant in the early days. The employment of CNN for RE was first presented in [70]. By concatenating the lexical and sentence level features extracted by CNN, the relationship between two marked nouns can be predicted through a softmax classifier. Nevertheless, simple CNN models often fail to identify critical cues, and many of them still require an external dependency parser. Based on this, some researchers [71,72] incorporated the attention model in CNN to capture both entity-specific attention and relation-specific pooling attention so as to detect more subtle cues about relation extraction. Apart from CNN, RNN-based models [73–75] are also employed for relation extraction. Analogously, attention mechanism [76–78] was introduced to capture the most important semantic information in a sentence. Benefiting from the superior performance of BERT [51], pretrained models are used for relation extraction [79,80]. However, supervised methods suffer from the expensive and limited labeled training data. In light of this, Mintz et al. [81] proposed a distant supervised approach, which intuitively extracts relations based on Freebase and named entity tagger. To alleviate the noise introduced by distant supervision, some researchers [82–85] resort to attention mechanism and GCN. In order to further avoid costly

data labeling and the wrong labeling problem caused by distant supervision methods, transfer learning and reinforcement learning have also recently been integrated into neural relation extraction. For transfer learning, Liu et al. [86] proposed a word-level distant supervised model initialized with prior knowledge learned from the relevant task of entity classification by transfer learning. Di et al. [87] proposed to explore a large amount of existing KBs that may not be very closely related to the target relation to extract relations mentioned within a given text corpus. For reinforcement learning(RL), Zeng et al. [88] proposed to learn sentence relations through the reinforcement learning method and with a distantly supervised dataset. To extract overlapping relations, Takanobu et al. [89] designed and incorporated reinforcement learning into an end-to-end hierarchical paradigm which decomposes the task into a relation detection task and an entity extraction task.

3.2.2. Multimodal Relation Extraction

Although the aforementioned RE methods have made considerable progress, there may be a performance decline in social media posts when texts lack context. Visual contents have been demonstrated to be effective in complementing textual contexts in other domains, such as named entity recognition, entity extraction and entity linking. Figure 5 shows an example of visual relation extraction, and with the support of visual information, the relation extraction can be more precise. In light of this, some methods use visual elements of documents to learn generalizable features that can be used in conjunction with textual semantics for better relation extraction. Zheng et al. [69] defined multimodal relation extraction(MRE) as the problem of classifying textual relations between two entities with visual information.

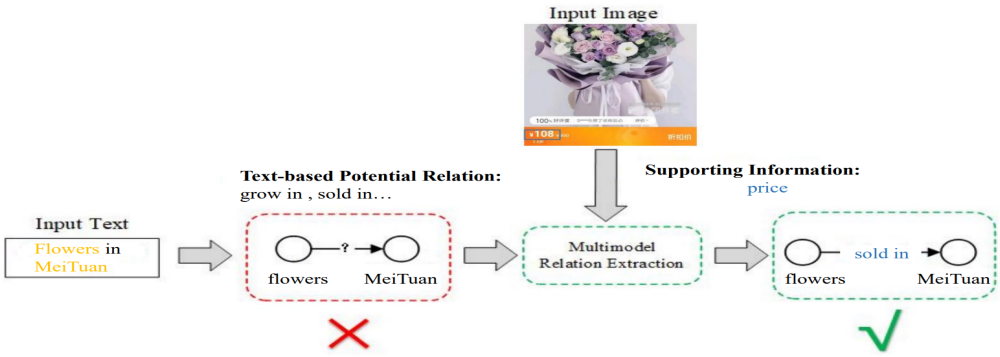


Figure 5. Illustration of visual relation extraction.

To address the lack of contexts for social media texts, Zheng et al. [90] firstly introduced a human-annotated multimodal dataset for testing the ability of neural relation extraction and proposed several multimodal baselines against previous SOTA text-based relation extraction models, showing that, through merging visual and textual information, the result for relation extraction can be significantly improved. Zheng et al. [69] further proposed a Multimodal Neural Network with Efficient Graph Alignment (MEGA), which utilizes graph-structured visual information to guide the extraction of textual relations with both semantic and structural graph similarity taking into consideration. To address high coupling in multimodal information and serious unbalanced distribution, Wan et al. [91] presented FL-MSRE, a few-shot learning-based approach, for extracting social relations with the help of both texts and face images. To alleviate the noise caused by irrelevant visual elements, Chen et al. [92] proposed MKGformer, which is a hybrid transformer for unified multimodal KGC. It models multimodal features of the entity across the last few layers of the visual transformer and the textual transformer with multilevel fusion. Chen et al. [93] proposed Hierarchical Visual Prefix fusion Network(HVPNeT), which generates effective and robust textual representation by incorporating hierarchical multiscaled visual features through a visual prefix-based attention mechanism at each self-attention layer of BERT for reducing error sensitivity.

3.2.3. Visual Relation Extraction and Grounding

We also introduce visual relation extraction and visual relation grounding: the former aims to extract the relation within image objects while the latter refers to discovering images, which contain or represent a specific relation, from an image corpus.

Visual Relation Extraction. Visual relation extraction plays a crucial role in the comprehensive understanding of an image through describing all the objects within the scene and how they relate to each other. Language priors is considered to provide helpful information to detect visual relationships. Lu et al. [94] fine-tuned the likelihood of the predicted relationship with the language. Inspired by the successful TransE method in knowledge graph representation learning, Zhang et al. [95] proposed VTransE, which measures embedding of predicates through mapping the visual features into the corresponding predicate space. Apart from language priors, statistical information is also utilized by some models for enhancing the performance. For example, Dai et al. [96] proposed to exploit the statistical dependencies among predicates, objects and subjects. Moreover, context messages passing also plays a vital role. For instance, through joint inference by iterative message passing, Xu et al. [97] were able to predict each visual relationship. However, the previous model focused on message passing in the same image, and Wang et al. [98] achieved messages passing from different images and were first to apply the one-shot learning approach to visual relationship detection.

Visual Relation Grounding. Visual relation grounding, along with visual grounding, has been widely studied and is now a subject under intense research. Mao et al. [99] were first to explore referring expression grounding by modeling images and sentences through CNN and LSTM. Grounding is achieved through extracting region proposals, along with finding the region which is able to generate the sentence possessing maximum posterior probability. Likewise, Rohrbach et al. [100] explored image grounding via reconstruction, which enables grounding in a weakly supervised manner. In order to explore referring relationships, Krishna et al. [101] utilized iterative message passing among subjects and objects. The above works mainly focus on image grounding. There are also some works focusing on video grounding. Zhou et al. [102] explored the weakly supervised grounding of descriptive nouns in distinct frames in a frame-weighted fashion. Huang et al. [103] investigated grounding referring expression in temporally instructional videos. Chen et al. [104] presented a model to perform spatio-temporal object grounding under video-level supervision. Specifically, they first pre-extracted the action tubes. Afterward, they ranked and returned the tube of maximum similarity using query sentences. Xiao et al. [105] were first to define the task of visual relation grounding in videos and proposed a weakly supervised approach for video relation grounding. Specifically, they collaboratively optimized two sequences of regions over a hierarchical spatio-temporal region graph and proposed a message passing mechanism based on spatial attention shifting between visual entities so as to pinpoint the related subject and objects.

3.3. Event Extraction

Event extraction (EE) is a long-standing and crucial task in information extraction research. As a particular form of information, an event refers to the occurrence of something at a specific time and place involving one or more participants, which can usually be described as a state change [106]. An event usually includes a trigger and several arguments with their corresponding argument roles. A trigger is usually a verb that marks the occurrence of the event; an argument is a word describing important information like time, place or participants; and an argument role refers to the role that an argument plays in the course of an event. Event extraction is intended to extract event information from unstructured texts, most of which illustrate the who, what, when, where, why and how of real-world events that have occurred. Nevertheless, the text-based EE models are generally limited due to the ambiguity of natural language. Events do not solely exist in a single modality of text, and similar event types, arguments or participants may coexist in multimedia content, such as videos and images. In light of this, researchers have proposed

multimedia event extraction, a task that aims to jointly extract events and arguments from multiple modalities.

This section provides a brief overview of text-based event extraction and a detailed summary of the development of multimedia event extraction thus far.

3.3.1. Text-Based Event Extraction

The traditional event extraction techniques are pattern-matching methods and statistical methods, with the latter achieving better results and becoming a research hotspot. The pattern-matching methods are mainly based on syntax trees or regular expressions, and their performance is strongly dependent on the expression form of text, domain, etc. Statistical learning to identify events, on the other hand, refers to the idea of text classification and transfers event detection and argument extraction into a classification problem, the core of which lies in the construction of classifiers and the selection of features. Nevertheless, it is challenging for traditional event extraction methods to learn in-depth features, making it difficult to improve the task of event extraction that depends on complex semantic relations. Compared with traditional event extraction methods, deep learning methods can capture complex semantic relations and significantly improve multiple event extraction datasets. Most recent event extraction works are based on deep learning architectures like CNN [107], RNN [108], transformer [109] and other networks [110,111].

3.3.2. Multimodal Event Extraction

Apart from textual modality, event types, arguments or participants may coexist in multimedia content [112]. Figure 6 illustrates that the visual information can correct the extraction result and enrich some event arguments that can only be extracted from the image. For this reason, researchers have proposed the multimedia event extraction task which leverages multimedia information.

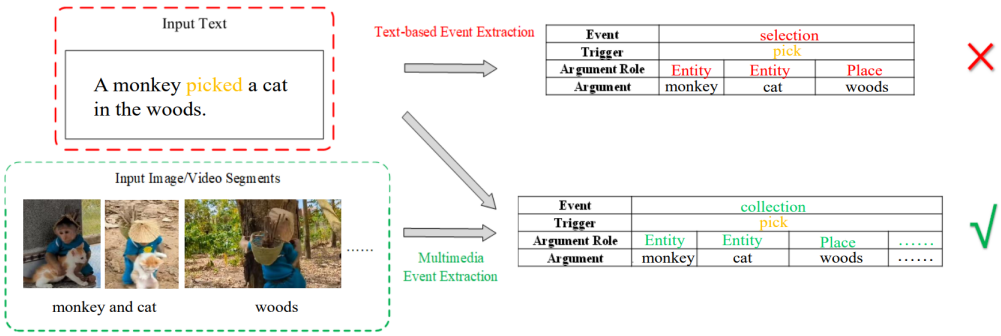


Figure 6. Illustration of multimodal event extraction.

Event Extraction with Text and Image. Image features are capable of providing additional information for event extraction. Zhang et al. [112] were first to propose a multimodal framework to integrate explicit visual information to resolve ambiguities of the text-only modality and improve event extraction performance on text documents. They adopted an in-domain visual pattern discovery method, which can be divided into visual argument discovery and visual feature extraction, to generate auxiliary background knowledge for each specific event. Ultimately, the visual features and text features extracted by JointIE [113] were integrated to improve event extraction performance. Li et al. [11] proposed the learning of a structured multimedia embedding space for multimedia event extraction. More specifically, they represented each image or sentence as a graph and adopted a weakly supervised framework to align the modalities. Moreover, they constructed an annotated news dataset called M^2E^2 as a benchmark for multimedia events extraction.

Event Extraction with Text and Video. Videos contain rich dynamics and detailed development of events, which can also be used as an information source for event detection. To this end, Chen et al. [114] introduced a new task of video multimedia event extraction

which extracts multimodal events and arguments from text and videos jointly. For this task, they proposed a self-supervised model to determine the coreference between video events and text events, and adopted the multimodal transformer to extract structured event information jointly from both videos and text documents. Analogously, the authors introduced a new multimodal video–text dataset with extensive annotations covering event and argument role extraction. Sadhu et al. [115] introduced visual semantic role labeling and proposed a framework for understanding and representing relevant salient events within videos. Specifically, they attempted to represent videos as a set of events, each of which is formed by a verb and several entities fulfilling various roles related to the event. Chen et al. [116] introduced the new task of vM^2E^2 and proposed two novel components to construct the first system of this task. Specifically, they proposed a self-supervised multimodal event coreference model that is able to determine event coreference between video and text events, as well as a multimodal transformer that jointly distills structured event information from videos and text documents.

4. Multimodal Knowledge Graph Completion

Knowledge graph completion has become an area of interest interest for various applications, as it involves completing the structure of a knowledge graph by predicting the missing entities or relationships within the knowledge graph as well as mining unknown facts. In this section, an extensive review of the research of multimodal knowledge graph completion is provided.

4.1. Entity Linking

Entity linking is an essential task in information extraction since it resolves the lexical ambiguity of entity mentions and determines their meanings in context. It refers to the task of linking entity mentions within texts with their corresponding entities in a knowledge base and one of its potential applications is knowledge base population [117], which means entity linking can be utilized for knowledge graph completion. Early approaches of entity linking focus on the language domain and generally lack visual information that can be used in this task, which hinders the construction of large-scale multimodal knowledge bases and poses great challenges for computing techniques to understand the real-world multimodal data comprehensively. To address this problem, an emerging task called multimodal entity linking was proposed, which utilizes both textual and visual information to map an ambiguous mention to an entity in a knowledge base [17].

This section provides a brief overview of text-based entity linking general architecture and a detailed summary of the multimodal entity linking approaches proposed thus far.

4.1.1. Text-Based Entity Linking

The universal approach to entity linking is to treat it as a ranking problem [118]. Figure 7 illustrates the generalized architecture, which is applicable to the majority of neural approaches.

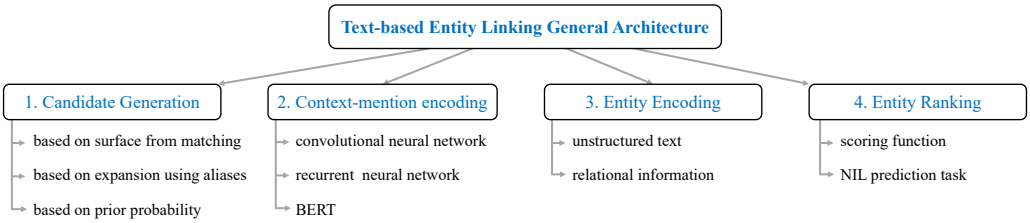


Figure 7. Reference graph of text-based entity linking general architecture.

An essential part of entity linking is candidate generation, which aims to perform preliminary filtering of the entity list. According to [118], there are three common candidate generation methods in entity linking: (1) those based on surface form matching, (2) those based on expansion using aliases and (3) those based on a prior probability computation.

The first approach generates a candidate list composed of entities that match diverse surface forms of mentions in the text [119–121]. The second approach constructs a dictionary of additional aliases using KG metadata like the disambiguation/redirect pages of Wikipedia to improve the candidate generation recall [122–124] while the third one is based on precalculated prior probabilities of correspondence between certain mentions and entities.

The next crucial step is context-mention encoding, which aims to correctly disambiguate an entity mention and capture the information from its context. The mainstream approach represents a mention as a dense contextualized vector. Early techniques mainly utilize a convolutional encoder as well as local neural attention to obtain the mention embedding [125–127]. In recent models, recurrent networks and self-attention mechanism prevail [128,129]. Additionally, encoding methods based on self-attention have recently become ubiquitous [130,131].

The third step is entity encoding, which aims to construct distributed vector representations of entity candidates. Some methods adopt entity representations constructed using relations between entities in knowledge graphs and graph embedding methods. Huang et al. [132] proposed a deep semantic relatedness model to generate dense entity representations from sparse entity features. Several methods expand their entity relatedness objective by aligning mentions and entities in a unified vector space [133–135]. As for graph embedding methods, representative works include DeepWalk [136] and TransE [22]. A few works [137,138] have verified the effectiveness of this method in entity linking. Apart from the aforementioned methods, recent work have utilized pretrained language models like BERT [51] for encoding entities [130,131].

The fourth step is entity ranking. Its goal is to give a list of entity candidates from a knowledge graph and a context with a mention to assign a score to each entity. During the calculation of the score, the mention representation is generated in the mention encoding step, and the entity representation is generated in the entity encoding step. Prevailing scoring functions for calculating similarity include dot product [127,128] and cosine similarity [125,139]. The final decision is deduced by a probability distribution, which can be approximated by a softmax function over the candidates.

It is important to note that since the corresponding entities of some mentions can be deficient in knowledge graphs, an entity linking system should be able to predict the lack of references if a mention appears in specific contexts, which is called NIL prediction task (NIl is the value returned when there is no corresponding entities during a entity linking task). Common ways to carry out the NIL prediction include setting a threshold for the best linking probability [140,141], introducing an additional "NIL" entity in the ranking stage [142] and training an additional binary classifier [120,143].

4.1.2. Multimodal Entity Linking

In addition to textual information, visual information is effective for depicting an entity auxiliary and is instrumental for the construction of large-scale multimodal knowledge bases. As Figure 8 shows, if we only take text into consideration, it is difficult for us to distinguish whether juustin should be linked to Justin Bieber or Justin Trudeau. However, with the help of some visual information, like the image in this example, which can be associated with concerts, it becomes obvious that juustin should be linked to Justin Bieber. To this end, researchers have proposed the multimodal entity linking task, which leverages both textual and visual information to disambiguate mentions by linking them to the corresponding entities in a knowledge base [17]. We first introduce multimodal entity linking datasets and then further classify multimodal entity linking approaches into multimodal attention-based approaches, which are jointly learning-based approaches and graph-based approaches.

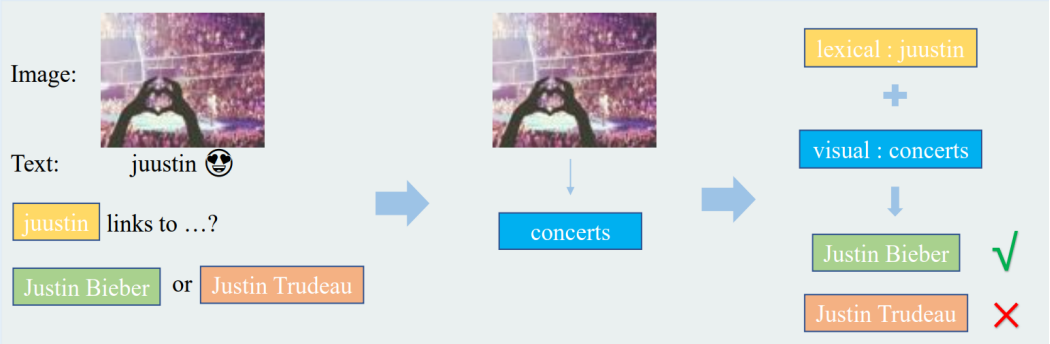


Figure 8. Illustration of multimodal entity linking.

Multimodal Entity Linking Datasets. The growing trend toward multimodality requires the expansion of EL research from single modality to multimodality. Multimodal Entity Linking (MEL) datasets have played a vital role in supporting multimodal entity linking tasking, and a wide range of research has been conducted on this topic. Moon et al. [144] were the first to address the MEL task and proposed their MEL dataset. However, their dataset is unavailable due to GDPR rules. To address this issue, Adjali et al. [13,145] proposed a framework to automatically build the MEL dataset from Twitter. However, this dataset has limited entity types along with ambiguous mentions, and thus it is not sufficiently challenging. Zhang et al. [17] proposed a Chinese MEL dataset collected from Weibo, which is a Chinese social media platform. However, their work mainly focuses on the person entities. Gan et al. [146] studied on a MEL dataset collected from movie reviews and proposed to disambiguate both visual and textual mentions. This dataset mostly focuses on the movie domain. Peng [146] released three MEL datasets, which were built from Weibo, Wikipedia and Richpedia information, with CNDBpedia, Wikidata and Richpedia being the respective, KBs. However, this dataset may lead to the data leakage problem, as many language models are pretrained on Wikipedia. To address the problem, Wang et al. [147] presented the WIKIDiverse dataset, which is a manually annotated high-quality MEL dataset covering diversified topics and entity types.

Multimodal Attention-based Methods. The attention mechanism has been proven to be superior for extracting the crucial parts from chaotic information. Zhang et al. [17] proposed an MEL model which can remove the impact of noisy images as well as better capture the interaction between mention representation and its corresponding entity representation with the help of a multiple attention mechanism. Moon et al. [144] built a zero-shot multimodal network and adopted a modality-attention module to attenuate irrelevant modalities while amplifying the most informative ones for entity disambiguation. Furthermore, they manually constructed a dataset called SnapCaptionsKB, which is composed of Snapchat image captions, with mentions fully annotated and linked to corresponding entities in an external knowledge base.

Joint Learning-based Methods. Some joint learning proposals are also applied for multimodal entity linking. Omar et al. [13] used joint learning of a representation of both mentions and entities based on the unsupervised Sent2Vec model [148] and pretrained the InceptionV3 model [64] for multimodal entity linking. In order to evaluate the multimodal entity linking model, they proposed a quasi-automatic dataset construction method and released a new annotated dataset of Twitter posts for this task.

Graph-Based Methods. Some works resort to a graph for addressing the MEL task. Gan et al. [146] formulated the entity linking task as a bipartite graph matching problem and proposed an optimal-transportation-based model to perform multimodal entity linking. Analogously, the authors released a finely labeled multimodal entity linking dataset, M3EL, that focuses on disambiguation of movie characters given textual documents and pictures. Differently from the aforementioned work which links the textual mentions to knowledge graphs, Zheng et al. [149] concentrated on visual entity linking which maps the visual objects detected from an image to the entities in the knowledge graph. Specifically, they

employed the scene graph for better visual scene understanding and utilized the GRU to extract textual features of objects from image caption. All features are aggregated to rank and map visual objects to the entities in the knowledge.

Knowledge Representation Learning. Knowledge representation learning, also known as knowledge graph embedding, has found important applications in miscellaneous entity-oriented tasks and quickly gained widespread attention [150]. The core idea is to learn the distributed representations of knowledge graphs by projecting entities and relations to low-dimensional dense vector spaces so as to simplify the manipulation while retaining the inherent structure of the knowledge graph. It has been largely applied in enormous downstream tasks including knowledge graph completion [22], semantic parsing [151] and relation extraction [152,153], among others. Previous works of knowledge representation learning mainly calculate the entity and relation embeddings based on structured triples, ignoring rich visual information of entities that intuitively describe the appearances and behaviors of this entity. To this end, Xie et al. [12] were first to attempt incorporating the multimodal information of entities for knowledge representation learning.

This section briefly presents text-based knowledge representation learning and representation learning in multimodal tasks, while also providing a summary of the multimodal knowledge representation learning approaches devised thus far.

4.1.3. Text-Based Knowledge Representation Learning

Previous text-based knowledge representation learning can be roughly classified into translation-based methods, semantic-matching-based methods and neural network-based methods. Translation-based methods calculate the distance between entities where the relation is regarded as the translation operation. They ensure the translation result of head entity vector $\boldsymbol{h}$ and relation vector $\boldsymbol{r}$ close to the tail entity vector $\boldsymbol{t}$. Primitive methods, like TransE [154], use the simple principle $\boldsymbol{h} + \boldsymbol{r} \approx \boldsymbol{t}$. Advanced models, like TransH [155] and KG2E [156], utilize more complex representation space and spatial transformation. Semantic-matching-based methods measure the plausibility of facts by matching the latent semantics of entities and relations embodied in their vector space representations. RESCAL [157] uses a bilinear formulation to calculate the score of a fact, DistMult [158] proposed simplified relation matrices, and ComplEx [159] extends DisMult to complex vector space to deal with antisymmetric relations. Neural network-based methods utilize deep models, such as CNN, RNN and transformer, to model the interactions between entities and relations . ConvE [160] and ConvKB [161] both adopt convolutional neural networks but reshaped entity vector and relation vector differently as model inputs. KG-BERT [162] borrowed the idea of the pretrained language model and used transformer to encode entities and relations.

4.1.4. Multimodal-Based Knowledge Representation Learning

Text-based knowledge representation learning models merely consider the single-modal structural information with triple facts, while multimodal information with rich semantics has been ignored. To utilize the multimodal information for knowledge representation learning of multimodal knowledge graphs, researchers began with utilizing multimodal representation learning modules to capture multimodal features and combined them with traditional knowledge representation learning measures. Before we summarize multimodal knowledge representation learning, we first review multimodal representation learning.

Multimodal Representation Learning. Research on encoding and using features from different modalities have been conducted for some time. Thus far, several deep learning methods, including probabilistic graphical models, multimodal autoencoders, attention mechanism and generative adversarial network, have been widely used in multimodal tasks. These approaches unify the representation of visual, audio or textual information, and enable neural networks to understand semantics across modalities more precisely.

Probabilistic Graphical Models-based Methods. Probabilistic models include deep belief networks (DBNs) [163] as well as deep Boltzmann machines (DBMs) [164]. Multimodal DBN is a typical example of probabilistic graphical models proposed by Srivastava and Salakhutdinov [165], which uses learned joint representation across modalities through shared restricted Boltzmann machine (RBM) hidden layer on top of image-specific and text-specific DBNs.

Multimodal Autoencoder-based Methods. Autoencoders have garnered much attention because of their ability to learn representations. Ngiam et al. [166] trained a bimodal deep autoencoder, which consists of two separate autoencoders for reconstructing both modalities when given only the visual data so as to discover correlations across different modalities. Similar to Ngiam's work, Silberer et al. [167] proposed a variant, which exploits stacked autoencoders to learn semantic representations integrating visual and textual information. Wang et al. [168] further proposed to impose orthogonal regularization on the weights so as to reduce the redundancy in the learned hashing representation. The above mentioned works learn multimodal representation within a common subspace, and some researchers tried to capture the correlation between different modalities. For example, Feng et al. [169] proposed a correspondence autoencoder mode, which is constructed by correlating hidden representations of two unimodal autoencoders. More specifically, the model first learns a set of independent but correlated representations for each single modality through modality-specific autoencoders and then minimizes the similarity loss between modalities to capture their correlation. Based on Feng et al.'s work, Wang et al. [170] designed a learning objective function which takes both the intermodal and intramodal important into consideration. Autoencoders can also be used for extracting intermediate features. Liu et al. [171] fused multiple modalities into a joint representation which contains intrinsic semantic information through stacked contractive autoencoders, and Hong et al. [172] constructed a hypergraph Laplacian with low-rank representation for multimodal fusion.

Attention Mechanism-based Methods. Attention mechanism allows the model to direct focus on certain parts of the inputs. Attention mechanism usually falls into the following categories: key-based attention and keyless attention. Key-based attention is useful for evaluating intramodality and intermodality importance, and has been widely exploited in visual description applications. When fusing multimodal features, the balance of the contribution for each modality is a key issue. Thus, several pieces of research [173–175] have been conducted and have shown that application performance can be improved through dynamically assigning weights to modality-specific features conditioned within contexts. Keyless attention is applied when the key is hard to obtain or define, and the attention mechanism is directly conducted on the localized features. Keyless attention is typically applied under classification or regression tasks. It is extremely suitable for multimodal feature fusion tasks, as it selects prominent cues from raw input, which has been proven in several studies [102,176,177]. In order to model the interactions between different modalities for a visual question-answering task, Lu et al. [178] proposed a coattention mechanism which jointly reasons about visual and question attention as well as a hierarchical architecture which coattends to the image and questions. Zedeh et al. [179] applied a multiattention mechanism to find distinct interactions between modalities. Zhou et al. [102] presented a model to fuse heterogeneous user behavior features through a a multiattention mechanism.

Generative Adversarial Network-based Methods. The generative adversarial network (GAN) is an emerging technique that has been widely applied and hosted with great success. Due to its unsupervised nature, GAN is rather effective for learning data representation without labels and has been extended to the multimodal representation learning field. For the text-to-image synthesis task, Reed et al. [180] proposed a GAN-based model for translating visual concepts from characters to pixels. Specifically, the model is composed of a generator network and a discriminator network: the former encodes the text input with noise, translating it into image, while the latter is used to determine whether the text and the image are compatible or not. Zhang et al. [181] proposed stacked

generative adversarial networks for text-to-image synthesis. Specifically, they decomposed the problem into a more manageable problem via a sketch-refinement process and introduced a conditioning augmentation technique which smooths the text encoding in the latent conditioning manifold. Reed et al. [182] took the object location information revealed by bounding boxes and key points into consideration and combined them with the text descriptions to learn what content should be drawn in which location. In cross-modal retrieval cases, Peng et al. [183] proposed a cross-modal GAN architecture which is able to explore intermodality and intramodality correlation simultaneously in generative and discriminative models: the former is formed through cross-modal convolutional autoencoders with weight-sharing constraint, while the the latter exploits two types of discriminative models to discriminate intramodality and intermodality at the same time. Xu et al. [184] used learning of cross-modal representations in a shared latent subspace. In order to place data of various modalities into a common hash space, Zhang et al. [185] preserved the manifold structure across different modalities. Based on CycleGAN [186], Wu et al. [187] proposed the learning of cross-modal hash functions without available paired training samples via cycle consistency loss.

Multimodal Knowledge Representation Learning. In recent years, entity-based methods have achieved excellent performance in multimodal knowledge representation learning by incorporating multimodal knowledge as new entities and relations into traditional knowledge representation learning methods. Xie et al. [12] applied earned knowledge representations with both triples and images. Specifically, they constructed an image encoder with an attention-based method so as to complete learning of the image-based representation for each entity as an extension of the structure-based representations learned from the traditional translation-based method. Similarly, Mousselly-Sergieh et al. [188] proposed a multimodal translation-based approach and designed a new type of margin-based ranking loss to incorporate the linguistic representations and visual representations for knowledge representation learning. In contrast, TransAE [189] combines a multimodal autoencoder and TransE [22] to jointly learn representation based on both the structural knowledge and multimodal knowledge of triplets. Semantic-matching-based methods and neural network-based methods have also been improved to fit the multimodal setting. Liu et al. [15] proposed PoE to find entity alignment within multimodal knowledge graphs via extracting relational, latent, numerical and visual features. Entity and embeddings are trained through the link prediction task. Pezeshkpour et al. [190] attempted to incorporate the multimodal information into existing relational models like DistMult [158] and ConvE [160] for knowledge representation. For this reason, they introduced domain-specific encoders to embed multimodal context, which can be used to score the truth value of a triple. Recently, Wang et al. [10] designed an RSME model, which is able to automatically enhance or decline the influence of visual context during the representation learning process, proving that the leveraging of visual information can help to generate better knowledge graph embeddings under appropriate circumstances.

5. Multimodal Knowledge Graph Applications

Multimodal knowledge graphs have gradually attracted the attention of researchers and have been applied in miscellaneous domains. In this section, we primarily summarize the applications of multimodal knowledge graphs.

5.1. Multimodal Knowledge Graphs in the Recommender System

A recommender system is a compelling information filtering system that aims to predict a customer's ratings or preferences for a product and has shown great potential to tackle information explosion problems, enhancing the user experience in various applications.

Knowledge graphs are widely used in recommender systems, which can alleviate data sparsity and cold start problems. However, most previous works ignore the importance of multimodal information like images and text descriptions. For example, before buying a product online, users tend to watch the images of this product and its descriptions as well

as the related user reviews about the product. Thus, it is important to introduce multimodal information into the knowledge graphs to better enhance recommender systems [18]. For this reason, Sun et al. [18] were first to introduce the multimodal knowledge graph into the recommender system and modeled the multimodal knowledge graph from two aspects: entity information aggregation and entity relation reasoning. Specifically, they employed a graph attention mechanism to conduct information propagation and a translational method to model the reasoning relation between entities. Extensive experiments have demonstrated that multimodal knowledge graphs can improve the quality of recommender systems.

5.2. Multimodal Knowledge Graphs in E-Commerce

Live-streaming sales are becoming increasingly popular in E-commerce, the core of which consists of encouraging customers to purchase in an online broadcasting room. In order to provide rich and attractive information about each product item, researchers have proposed leveraging the multimodal knowledge graphs to enable customers to better understand a product without jumping out. In light of this, Xu et al. [19] constructed a multimodal knowledge graph called AliMe MKG that centers on and aggregates rich information about items. They further built an online live assistant based on AliMe MKG for product search, product exhibition and question answering, allowing customers to conveniently seek information in an online broadcasting room. Furthermore, AliMe MKG can be applied in short video productions which present the core selling points of a product item in an attractive and intuitive manner.

5.3. Multimodal Knowledge Graphs in Biomedicine

The knowledge graph is becoming increasingly significant in the biomedical field given the exponential increase in the volume of biomedical articles which contain valuable knowledge regarding biomedical entities such as proteins and drugs. Since protein–protein interaction(PPI) is one of the most important tasks in biomedical document processing, where the relation ("interaction" or "noninteraction") is determined from the given biomedical texts, and the knowledge of protein interaction is vital for understanding the biological processes, there have been some works on PPI in the literature. However, earlier works mostly focused on textual information within the biomedical texts, which lacks the ability to capture multiomics information related to protein interaction or the genetic and structural information of proteins. Toward this end, Pingali et al. [191] explored the Graph-based Transformer model (GraphBERT) [192] so as to learn the modality-independent graph representation. Specifically, they utilized protein atomic structural information when identifying the protein interactions and developed a generalized modality-agnostic approach that is capable of learning the feature representations for both textual and protein structural modalities.

5.4. Multimodal Knowledge Graphs in Fake News Detection

The widespread dissemination and misleading effect of online rumors on social media have become a key concern for the public and government. Discovering and regulating social media rumors is important to ensuring that users receive truthful information and maintain social harmony. Previous related work has mainly focused on inferring clues from the media content and social context, largely ignoring the rich knowledge information behind the highly condensed text, which is useful for rumor verification. Additionally, earlier methods perform badly on unseen events, as they tend to capture a plethora of event-specific features within seen data that cannot be transferred to newly emerging events. To address these problems, Zhang et al. [193] proposed Multi-modal Knowledge-aware Event Memory Network (MKEMN), which exploits the Multi-modal Knowledge-aware Network (MKN) and Event Memory Network (EMN) as building blocks for social media rumor detection. Specifically, MKN learns multimodal representations of posts on social media and retrieves external knowledge from real-world knowledge graphs so as to complement the semantic representation for short text in posts. Conceptual knowledge

is considered to be the additional evidence for improved rumor detection. Apart from the above work, Wang et al. [194] designed Knowledge-driven Multimodal Graph Convolutional Networks (KMGCNs) to model the semantic representation through modeling the textual information, knowledge concepts and visual information jointly into a unified framework for fake news detection. To this end, they converted each post into a graph and proposed a multimodal graph convolution network for capturing nonconsecutive phrases for better obtaining the composition of semantics. In addition, knowledge distillation is exploited for providing complementary knowledge concepts, which have better generalization performance for emerging posts.

6. Conclusions

With the proliferation of multimodal resources in the Internet and the development of related research in recent years, the subject of multimodal knowledge graphs has attracted increased attention. This paper reviews the related advances of multimodal knowledge graphs in the following three areas: open multimodal knowledge graphs, multimodal knowledge graph construction and completion technologies and typical applications. The related construction technologies are further elaborated from the three perspectives of named entity recognition, relation extraction and event extraction. Finally, we enumerate the typical applications of multimodal knowledge in various scenarios, such as recommender system and E-commerce. We hope that this survey can provide a good reference in facilitating future research.

Author Contributions: Conceptualization, Y.C.; Investigation, Y.C., X.G. and S.Y.; Resources, X.G. and L.H.; Writing—original draft, Y.C. and J.Z.; Writing—review & editing, X.G., S.Y., L.H. and J.L.; Supervision, X.G. All authors have read and agreed to the published version of the manuscript.

Funding: This work is supported by Beijing Academy of Artificial Intelligence.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Klyne, G.; Carroll, J.J. Resource Description Framework (RDF): Concepts and Abstract Syntax—W3C Recommendation 10 February 2004. Available online: https://www.w3.org/TR/rdf-concepts/ (accessed on 2 March 2023).
2. Zhang, F.; Yuan, N.J.; Lian, D.; Xie, X.; Ma, W. Collaborative Knowledge Base Embedding for Recommender Systems. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; pp. 353–362.
3. Wu, S.; Li, Y.; Zhang, D.; Zhou, Y.; Wu, Z. Diverse and Informative Dialogue Generation with Context-Specific Commonsense Knowledge Awareness. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online, 5–10 July 2020; pp. 5811–5820.
4. Yih, W.; Chang, M.; He, X.; Gao, J. Semantic Parsing via Staged Query Graph Generation: Question Answering with Knowledge Base. In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, Beijing, China, 26–31 July 2015; pp. 1321–1331.
5. Ilievski, F.; Szekely, P.; Zhang, B. Cskg: The commonsense knowledge graph. In Proceedings of the Semantic Web: 18th International Conference, ESWC 2021, Virtual Event, 6–10 June 2021; pp. 680–696.
6. Bollacker, K.D.; Cook, R.P.; Tufts, P. Freebase: A Shared Database of Structured General Human Knowledge. In Proceedings of the Twenty-Second AAAI Conference on Artificial Intelligence, Vancouver, BC, Canada, 22–26 July 2007; pp. 1962–1963.
7. Auer, S.; Bizer, C.; Kobilarov, G.; Lehmann, J.; Cyganiak, R.; Ives, Z.G. DBpedia: A Nucleus for a Web of Open Data. In *Semantic Web, 6th International Semantic Web Conference, 2nd Asian Semantic Web Conference, Busan, Republic of Korea, 11–15 November 2007*; Springer: Cham, Switzerland, 2007; pp. 722–735.
8. Vrandečić, D.; Krötzsch, M. Wikidata: A Free Collaborative Knowledgebase. *Commun. ACM* **2014**, 57 , 78–85. [CrossRef]
9. McCray, A.T. An upper-level ontology for the biomedical domain. *Comp. Funct. Genom.* **2003**, *4*, 80–84. [CrossRef] [PubMed]
10. Wang, M.; Wang, S.; Yang, H.; Zhang, Z.; Chen, X.; Qi, G. Is Visual Context Really Helpful for Knowledge Graph? A Representation Learning Perspective. In Proceedings of the 29th ACM International Conference on Multimedia, Chengdu, China, 20–24 October 2021; pp. 2735–2743.

11. Li, M.; Zareian, A.; Zeng, Q.; Whitehead, S.; Lu, D.; Ji, H.; Chang, S. Cross-media Structured Common Space for Multimedia Event Extraction. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online, 5–10 July 2020; pp. 2557–2568.
12. Xie, R.; Liu, Z.; Luan, H.; Sun, M. Image-embodied Knowledge Representation Learning. In Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, Melbourne, Australia, 19–25 August 2017; pp. 3140–3146.
13. Adjali, O.; Besançon, R.; Ferret, O.; Borgne, H.L.; Grau, B. Multimodal Entity Linking for Tweets. In *Advances in Information Retrieval—42nd European Conference on IR Research, Lisbon, Portugal, 14–17 April 2020*; Springer: Cham, Switzerland, 2020; pp. 463–478.
14. Ferrada, S.; Bustos, B.; Hogan, A. IMGpedia: A Linked Dataset with Content-Based Analysis of Wikimedia Images. In *Semantic Web—ISWC 2017—16th International Semantic Web Conference, Vienna, Austria, 21–25 October 2017*; Springer: Cham, Switzerland, 2017; pp. 84–93.
15. Liu, Y.; Li, H.; García-Durán, A.; Niepert, M.; Oñoro-Rubio, D.; Rosenblum, D.S. MMKG: Multi-modal Knowledge Graphs. In *Semantic Web—16th International Conference, Portorož, Slovenia, 2–6 June 2019*; Springer: Cham, Switzerland, 2019; pp. 459–474.
16. Wang, M.; Wang, H.; Qi, G.; Zheng, Q. Richpedia: A Large-Scale, Comprehensive Multi-Modal Knowledge Graph. *Big Data Res.* **2020**, *22* 100159. [CrossRef]
17. Zhang, L.; Li, Z.; Yang, Q. Attention-Based Multimodal Entity Linking with High-Quality Images. In *Database Systems for Advanced Applications—26th International Conference, Taipei, Taiwan, 11–14 April 2021*; Springer: Cham, Switzerland, 2021; Volume 12682, pp. 533–548.
18. Sun, R.; Cao, X.; Zhao, Y.; Wan, J.; Zhou, K.; Zhang, F.; Wang, Z.; Zheng, K. Multi-modal Knowledge Graphs for Recommender Systems. In Proceedings of the CIKM '20: The 29th ACM International Conference on Information and Knowledge Management, Virtual Event, 19–23 October 2020; pp. 1405–1414.
19. Xu, G.; Chen, H.; Li, F.; Sun, F.; Shi, Y.; Zeng, Z.; Zhou, W.; Zhao, Z.; Zhang, J. AliMe MKG: A Multi-modal Knowledge Graph for Live-streaming E-commerce. In Proceedings of the CIKM '21: The 30th ACM International Conference on Information and Knowledge Management, Gold Coast, Australia, 1–5 November 2021 ; pp. 4808–4812.
20. Zhu, X.; Li, Z.; Wang, X.; Jiang, X.; Sun, P.; Wang, X.; Xiao, Y.; Yuan, N.J. Multi-Modal Knowledge Graph Construction and Application: A Survey. *arXiv* **2022**, arXiv:2202.05786.
21. Toutanova, K.; Chen, D. Observed Versus Latent Features for Knowledge Base and Text Inference. In Proceedings of the 3rd Workshop on Continuous Vector Space Models and Their Compositionality, Beijing, China, 31 July 2015.
22. Bordes, A.; Usunier, N.; García-Durán, A.; Weston, J.; Yakhnenko, O. Translating Embeddings for Modeling Multi-relational Data. In Proceedings of the Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems, Lake Tahoe, NV, USA, 5–10 December 2013 2013; pp. 2787–2795.
23. Suchanek, F.M.; Kasneci, G.; Weikum, G. Yago: A core of semantic knowledge. In Proceedings of the 16th International Conference on World Wide Web, Banff, AB, Canada, 8–12 May 2007; pp. 697–706.
24. Oñoro-Rubio, D.; Niepert, M.; García-Durán, A.; Gonzalez-Sanchez, R.; López-Sastre, R.J. Answering Visual-Relational Queries in Web-Extracted Knowledge Graphs. In Proceedings of the Conference on Automated Knowledge Base Construction, Long Beach, CA, USA, 8 December 2017.
25. Alberts, H.; Huang, N.; Deshpande, Y.; Liu, Y.; Cho, K.; Vania, C.; Calixto, I. VisualSem: A high-quality knowledge graph for vision and language. In Proceedings of the 1st Workshop on Multilingual Representation Learning, Punta Cana, Dominican Republic, 7–9 November 2021; pp. 138–152. [CrossRef]
26. Navigli, R.; Ponzetto, S.P. BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artif. Intell.* **2012**, *193*, 217–250. [CrossRef]
27. Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.S.; et al. ImageNet Large Scale Visual Recognition Challenge. *Int. J. Comput. Vis.* **2014**, *115*, 211–252. [CrossRef]
28. Radford, A.; Kim, J.W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. Learning Transferable Visual Models From Natural Language Supervision. In Proceedings of the International Conference on Machine Learning, Virtual, 18–24 July 2021.
29. Li, J.; Sun, A.; Han, J.; Li, C. A Survey on Deep Learning for Named Entity Recognition. *IEEE Trans. Knowl. Data Eng.* **2022**, *34*, 50–70. [CrossRef]
30. Guo, J.; Xu, G.; Cheng, X.; Li, H. Named entity recognition in query. In Proceedings of the 32nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Boston, MA, USA, 19–23 July 2009; pp. 267–274.
31. Petkova, D.; Croft, W.B. Proximity-based document representation for named entity retrieval. In Proceedings of the Sixteenth ACM Conference on Information and Knowledge Management, Lisbon, Portugal, 6–10 November 2007; pp. 731–740.
32. Etzioni, O.; Cafarella, M.J.; Downey, D.; Popescu, A.; Shaked, T.; Soderland, S.; Weld, D.S.; Yates, A. Unsupervised named-entity extraction from the Web: An experimental study. *Artif. Intell.* **2005**, *165*, 91–134. [CrossRef]
33. Aliod, D.M.; van Zaanen, M.; Smith, D. Named Entity Recognition for Question Answering. In Proceedings of the Australasian Language Technology Workshop 2006, Sydney, Australia, November 2006; pp. 51–58.
34. Babych, B.; Hartley, A. Improving Machine Translation Quality with Automatic Named Entity Recognition. In Proceedings of the 7th International EAMT Workshop on MT and Other Language Technology Tools, Improving MT through Other Language Technology Tools, Resource and Tools for Building, Budapest, Hungary, 13 April 2003.

35. Humphreys, K.; Gaizauskas, R.J.; Azzam, S.; Huyck, C.; Mitchell, B.; Cunningham, H.; Wilks, Y. University of Sheffield: Description of the LaSIE-II System as Used for MUC-7. In Proceedings of the Seventh Message Understanding Conference: Proceedings of a Conference, Fairfax, VA, USA, 29 April–1 May 1998.
36. Aone, C.; Halverson, L.; Hampton, T.; Ramos-Santacruz, M. SRA: Description of the IE2 System Used for MUC-7. In Proceedings of the Seventh Message Understanding Conference: Proceedings of a Conference, Fairfax, VA, USA, 29 April–1 May 1998.
37. Appelt, D.E.; Hobbs, J.R.; Bear, J.; Israel, D.J.; Kameyama, M.; Martin, D.L.; Myers, K.L.; Tyson, M. SRI International FASTUS system: MUC-6 test results and analysis. In Proceedings of the 6th Conference on Message Understanding, Columbia, MD, USA, 6–8 November 1995; pp. 237–248.
38. Mikheev, A.; Moens, M.; Grover, C. Named Entity Recognition without Gazetteers. In Proceedings of the EACL 1999, 9th Conference of the European Chapter of the Association for Computational Linguistics. The Association for Computer Linguistics, Bergen, Norway, 8–12 June 1999; pp. 1–8.
39. Bikel, D.M.; Miller, S.; Schwartz, R.M.; Weischedel, R.M. Nymble: A High-Performance Learning Name-finder. In Proceedings of the 5th Applied Natural Language Processing Conference, Washington, DC, USA, 31 March–3 April 1997; pp. 194–201.
40. Bikel, D.M.; Schwartz, R.M.; Weischedel, R.M. An Algorithm that Learns What's in a Name. *Mach. Learn.* **1999**, *34*, 211–231. [CrossRef]
41. Quinlan, J.R. Induction of Decision Trees. *Mach. Learn.* **1986**, *1*, 81–106. [CrossRef]
42. Szarvas, G.; Farkas, R.; Kocsor, A. A Multilingual Named Entity Recognition System Using Boosting and C4.5 Decision Tree Learning Algorithms. In *Discovery Science, 9th International Conference, Barcelona, Spain, 7–10 October 2006*; Springer: Cham, Switzerland, 2006; pp. 267–278.
43. Lafferty, J.D.; McCallum, A.; Pereira, F.C.N. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In Proceedings of the Eighteenth International Conference on Machine Learning (ICML 2001), Williamstown, MA, USA, 28 June–1 July 2001; pp. 282–289.
44. McCallum, A.; Li, W. Early results for Named Entity Recognition with Conditional Random Fields, Feature Induction and Web-Enhanced Lexicons. In Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003, CoNLL 2003, Edmonton, AB, Canada, 31 May 2003; pp. 188–191.
45. Krishnan, V.; Manning, C.D. An Effective Two-Stage Model for Exploiting Non-Local Dependencies in Named Entity Recognition. In Proceedings of the ACL 2006, 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics, Sydney, Australia, 17–18 July 2006.
46. Wu, Y.; Jiang, M.; Lei, J.; Xu, H. Named Entity Recognition in Chinese Clinical Text Using Deep Neural Network. In *MEDINFO 2015: eHealth-enabled Health— Proceedings of the 15th World Congress on Health and Biomedical Informatics, São Paulo, Brazil, 19–23 August 2015*; IOS Press: Amsterdam, The Netherlands, 2015; pp. 624–628.
47. Collobert, R.; Weston, J.; Bottou, L.; Karlen, M.; Kavukcuoglu, K.; Kuksa, P.P. Natural Language Processing (Almost) from Scratch. *J. Mach. Learn. Res.* **2011**, *12*, 2493–2537.
48. Strubell, E.; Verga, P.; Belanger, D.; McCallum, A. Fast and Accurate Entity Recognition with Iterated Dilated Convolutions. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Copenhagen, Denmark, 7–11 September 2017; pp. 2670–2680.
49. Huang, Z.; Xu, W.; Yu, K. Bidirectional LSTM-CRF Models for Sequence Tagging. *arXiv* **2015**, arXiv:1508.01991.
50. Zhang, Y.; Yang, J. Chinese NER Using Lattice LSTM. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, Melbourne, Australia, 15–20 July 2018; pp. 1554–1564.
51. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Minneapolis, MN, USA, 2–7 June 2019; pp. 4171–4186.
52. Zhang, T.; Xia, C.; Yu, P.S.; Liu, Z.; Zhao, S. PDALN: Progressive Domain Adaptation over a Pre-trained Model for Low-Resource Cross-Domain Named Entity Recognition. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Punta Cana, Dominican Republic, 7–11 November 2021; pp. 5441–5451.
53. Liu, J.; Gao, L.; Guo, S.; Ding, R.; Huang, X.; Ye, L.; Meng, Q.; Nazari, A.; Thiruvady, D. A hybrid deep-learning approach for complex biochemical named entity recognition. *Knowl. Based Syst.* **2021**, *221*, 106958. [CrossRef]
54. Fang, Z.; Zhang, Q.; Kok, S.; Li, L.; Wang, A.; Yang, S. Referent graph embedding model for name entity recognition of Chinese car reviews. *Knowl. Based Syst.* **2021**, *233*, 107558. [CrossRef]
55. Moon, S.; Neves, L.; Carvalho, V. Multimodal Named Entity Recognition for Short Social Media Posts. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, New Orleans, LA, USA, 1–6 June 2018; pp. 852–860.
56. Zhang, Q.; Fu, J.; Liu, X.; Huang, X. Adaptive Co-attention Network for Named Entity Recognition in Tweets. In Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018; pp. 5674–5681.
57. Shahzad, M.; Amin, A.; Esteves, D.; Ngomo, A.N. InferNER: An attentive model leveraging the sentence-level information for Named Entity Recognition in Microblogs. In Proceedings of the Thirty-Fourth International Florida Artificial Intelligence Research Society Conference, North Miami Beach, FL, USA, 17–19 May 2021.

58. Lu, D.; Neves, L.; Carvalho, V.; Zhang, N.; Ji, H. Visual Attention Model for Name Tagging in Multimodal Social Media. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, Melbourne, Australia, 15–20 July 2018; pp. 1990–1999.
59. Arshad, O.; Gallo, I.; Nawaz, S.; Calefati, A. Aiding Intra-Text Representations with Visual Context for Multimodal Named Entity Recognition. In Proceedings of the 2019 International Conference on Document Analysis and Recognition, Sydney, Australia, 20–25 September 2019; pp. 337–342.
60. Wu, Z.; Zheng, C.; Cai, Y.; Chen, J.; Leung, H.; Li, Q. Multimodal Representation with Embedded Visual Guiding Objects for Named Entity Recognition in Social Media Posts. In Proceedings of the MM '20: The 28th ACM International Conference on Multimedia, Seattle, WA, USA, 12–16 October 2020; pp. 1038–1046.
61. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R.B. Mask R-CNN. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2980–2988.
62. Zheng, C.; Wu, Z.; Wang, T.; Cai, Y.; Li, Q. Object-Aware Multimodal Named Entity Recognition in Social Media Posts With Adversarial Learning. *IEEE Trans. Multim.* **2021**, *23*, 2520–2532. [CrossRef]
63. Asgari-Chenaghlu, M.; Feizi-Derakhshi, M.; Farzinvash, L.; Motamed, C. A multimodal deep learning approach for named entity recognition from social media. *arXiv* **2020**, arXiv:2001.06888.
64. Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; Wojna, Z. Rethinking the Inception Architecture for Computer Vision. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 2818–2826.
65. Sun, L.; Wang, J.; Su, Y.; Weng, F.; Sun, Y.; Zheng, Z.; Chen, Y. RIVA: A Pre-trained Tweet Multimodal Model Based on Text-image Relation for Multimodal NER. In Proceedings of the 28th International Conference on Computational Linguistics, Barcelona, Spain, 8–13 December 2020; pp. 1852–1862.
66. Sun, L.; Wang, J.; Zhang, K.; Su, Y.; Weng, F. RpBERT: A Text-image Relation Propagation-based BERT Model for Multimodal NER. In Proceedings of the Thirty-Fifth AAAI Conference on Artificial Intelligence, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, The Eleventh Symposium on Educational Advances in Artificial Intelligene, Virtual Event, 2–9 February 2021; pp. 13860–13868.
67. Yu, J.; Jiang, J.; Yang, L.; Xia, R. Improving Multimodal Named Entity Recognition via Entity Span Detection with Unified Multimodal Transformer. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online, 5–10 July 2020; pp. 3342–3352.
68. Zhang, D.; Wei, S.; Li, S.; Wu, H.; Zhu, Q.; Zhou, G. Multi-modal Graph Fusion for Named Entity Recognition with Targeted Visual Guidance. In Proceedings of the Thirty-Fifth AAAI Conference on Artificial Intelligence, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, The Eleventh Symposium on Educational Advances in Artificial Intelligence, Virtual Event, 2–9 February 2021; pp. 14347–14355.
69. Zheng, C.; Feng, J.; Fu, Z.; Cai, Y.; Li, Q.; Wang, T. Multimodal Relation Extraction with Efficient Graph Alignment. In Proceedings of the MM '21: ACM Multimedia Conference, Virtual Event, 20–24 October 2021; pp. 5298–5306.
70. Zeng, D.; Liu, K.; Lai, S.; Zhou, G.; Zhao, J. Relation Classification via Convolutional Deep Neural Network. In Proceedings of the COLING 2014, 25th International Conference on Computational Linguistics, Dublin, Ireland, 23–29 August 2014; pp. 2335–2344.
71. Shen, Y.; Huang, X. Attention-Based Convolutional Neural Network for Semantic Relation Extraction. In Proceedings of the COLING 2016, 26th International Conference on Computational Linguistics, Osaka, Japan, 11–16 December 2016; pp. 2526–2536.
72. Wang, L.; Cao, Z.; de Melo, G.; Liu, Z. Relation Classification via Multi-Level Attention CNNs. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, Berlin, Germany, 7–12 August 2016.
73. Miwa, M.; Bansal, M. End-to-End Relation Extraction using LSTMs on Sequences and Tree Structures. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, Berlin, Germany, 7–12 August 2016.
74. Zhang, S.; Zheng, D.; Hu, X.; Yang, M. Bidirectional Long Short-Term Memory Networks for Relation Classification. In Proceedings of the 29th Pacific Asia Conference on Language, Information and Computation, Shanghai, China, 30 October–1 November 2015.
75. Xu, Y.; Mou, L.; Li, G.; Chen, Y.; Peng, H.; Jin, Z. Classifying Relations via Long Short Term Memory Networks along Shortest Dependency Paths. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP, Lisbon, Portugal, 17–21 September 2015; pp. 1785–1794.
76. Zhou, P.; Shi, W.; Tian, J.; Qi, Z.; Li, B.; Hao, H.; Xu, B. Attention-Based Bidirectional Long Short-Term Memory Networks for Relation Classification. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, Berlin, Germany, 7–12 August 2016.
77. Xiao, M.; Liu, C. Semantic Relation Classification via Hierarchical Recurrent Neural Network with Attention. In Proceedings of the COLING 2016, 26th International Conference on Computational Linguistics, Osaka, Japan, 11–16 December 2016; pp. 1254–1263.
78. Lee, J.; Seo, S.; Choi, Y.S. Semantic Relation Classification via Bidirectional LSTM Networks with Entity-Aware Attention Using Latent Entity Typing. *Symmetry* **2019**, *11*, 785. [CrossRef]
79. Wu, S.; He, Y. Enriching Pre-trained Language Model with Entity Information for Relation Classification. In Proceedings of the 28th ACM International Conference on Information and Knowledge Management, Beijing, China, 3–7 November 2019; pp. 2361–2364.

80. Yamada, I.; Asai, A.; Shindo, H.; Takeda, H.; Matsumoto, Y. LUKE: Deep Contextualized Entity Representations with Entity-aware Self-attention. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, Online, 16–20 November 2020; pp. 6442–6454.
81. Mintz, M.; Bills, S.; Snow, R.; Jurafsky, D. Distant supervision for relation extraction without labeled data. In Proceedings of the ACL 2009, Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the AFNLP, Singapore, 2–7 August 2009; pp. 1003–1011.
82. Han, X.; Liu, Z.; Sun, M. Neural Knowledge Acquisition via Mutual Attention between Knowledge Graph and Text. In Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018; AAAI'18/IAAI'18/EAAI'18.
83. Zeng, D.; Liu, K.; Chen, Y.; Zhao, J. Distant Supervision for Relation Extraction via Piecewise Convolutional Neural Networks. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, Lisbon, Portugal, 17–21 September 2015; pp. 1753–1762. [CrossRef]
84. Ye, Z.X.; Ling, Z.H. Distant Supervision Relation Extraction with Intra-Bag and Inter-Bag Attentions. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Minneapolis, MN, USA, 2–7 June 2019; Volume 1, pp. 2810–2819. [CrossRef]
85. Huang, W.; Mao, Y.; Yang, L.; Yang, Z.; Long, J. Local-to-global GCN with knowledge-aware representation for distantly supervised relation extraction. *Knowl. Based Syst.* **2021**, *234*, 107565. [CrossRef]
86. Liu, T.; Zhang, X.; Zhou, W.; Jia, W. Neural Relation Extraction via Inner-Sentence Noise Reduction and Transfer Learning. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, 31 October–4 November 2018; pp. 2195–2204.
87. Di, S.; Shen, Y.; Chen, L. Relation Extraction via Domain-aware Transfer Learning. In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, Anchorage, AK, USA, 4–8 August 2019; pp. 1348–1357.
88. Zeng, X.; He, S.; Liu, K.; Zhao, J. Large Scaled Relation Extraction With Reinforcement Learning. In Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, the 30th innovative Applications of Artificial Intelligence, and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018; pp. 5658–5665.
89. Takanobu, R.; Zhang, T.; Liu, J.; Huang, M. A Hierarchical Framework for Relation Extraction with Reinforcement Learning. In Proceedings of the The Thirty-Third AAAI Conference on Artificial Intelligence, The Thirty-First Innovative Applications of Artificial Intelligence Conference, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, Honolulu, HI, USA, 27 January–1 February 2019; pp. 7072–7079.
90. Zheng, C.; Wu, Z.; Feng, J.; Fu, Z.; Cai, Y. MNRE: A Challenge Multimodal Dataset for Neural Relation Extraction with Visual Evidence in Social Media Posts. In Proceedings of the 2021 IEEE International Conference on Multimedia and Expo (ICME), Shenzhen, China, 5–9 July 2021; pp. 1–6.
91. Wan, H.; Zhang, M.; Du, J.; Huang, Z.; Yang, Y.; Pan, J.Z. FL-MSRE: A few-shot learning based approach to multimodal social relation extraction. In Proceedings of the AAAI Conference on Artificial Intelligence, Virtually, 2–9 February 2021; Volume 35, pp. 13916–13923.
92. Chen, X.; Zhang, N.; Li, L.; Deng, S.; Tan, C.; Xu, C.; Huang, F.; Si, L.; Chen, H. Hybrid Transformer with Multi-level Fusion for Multimodal Knowledge Graph Completion. *arXiv* **2022**, arXiv:2205.02357.
93. Chen, X.; Zhang, N.; Li, L.; Yao, Y.; Deng, S.; Tan, C.; Huang, F.; Si, L.; Chen, H. Good Visual Guidance Makes A Better Extractor: Hierarchical Visual Prefix for Multimodal Entity and Relation Extraction. *arXiv* **2022**, arXiv:2205.03521.
94. Lu, C.; Krishna, R.; Bernstein, M.; Fei-Fei, L. Visual relationship detection with language priors. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016; pp. 852–869.
95. Zhang, H.; Kyaw, Z.; Chang, S.F.; Chua, T.S. Visual translation embedding network for visual relation detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 5532–5540.
96. Dai, B.; Zhang, Y.; Lin, D. Detecting visual relationships with deep relational networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 3076–3086.
97. Xu, D.; Zhu, Y.; Choy, C.B.; Fei-Fei, L. Scene graph generation by iterative message passing. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 5410–5419.
98. Wang, W.; Wang, M.; Wang, S.; Long, G.; Yao, L.; Qi, G.; Chen, Y. One-shot learning for long-tail visual relation detection. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; Volume 34, pp. 12225–12232.
99. Mao, J.; Huang, J.; Toshev, A.; Camburu, O.; Yuille, A.L.; Murphy, K. Generation and comprehension of unambiguous object descriptions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 11–20.
100. Rohrbach, A.; Rohrbach, M.; Hu, R.; Darrell, T.; Schiele, B. Grounding of textual phrases in images by reconstruction. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016; pp. 817–834.
101. Krishna, R.; Chami, I.; Bernstein, M.; Fei-Fei, L. Referring relationships. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 6867–6876.

102. Zhou, C.; Bai, J.; Song, J.; Liu, X.; Zhao, Z.; Chen, X.; Gao, J. Atrank: An attention-based user behavior modeling framework for recommendation. In Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018.
103. Huang, D.A.; Buch, S.; Dery, L.; Garg, A.; Fei-Fei, L.; Niebles, J.C. Finding "it": Weakly-supervised reference-aware visual grounding in instructional videos. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 5948–5957.
104. Chen, Z.; Ma, L.; Luo, W.; Wong, K.Y.K. Weakly-supervised spatio-temporally grounding natural sentence in video. *arXiv* **2019**, arXiv:1906.02549.
105. Xiao, J.; Shang, X.; Yang, X.; Tang, S.; Chua, T.S. Visual relation grounding in videos. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; pp. 447–464.
106. Doddington, G.R.; Mitchell, A.; Przybocki, M.A.; Ramshaw, L.A.; Strassel, S.M.; Weischedel, R.M. The Automatic Content Extraction (ACE) Program—Tasks, Data, and Evaluation. In Proceedings of the Fourth International Conference on Language Resources and Evaluation, Lisbon, Portugal, 26–28 May 2004.
107. Chen, Y.; Xu, L.; Liu, K.; Zeng, D.; Zhao, J. Event Extraction via Dynamic Multi-Pooling Convolutional Neural Networks. In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing. The Association for Computer Linguistics, Beijing, China, 26–31 July 2015; pp. 167–176.
108. Nguyen, T.H.; Cho, K.; Grishman, R. Joint Event Extraction via Recurrent Neural Networks. In Proceedings of the NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego, CA, USA, 12–17 June 2016; pp. 300–309.
109. Lv, J.; Zhang, Z.; Jin, L.; Li, S.; Li, X.; Xu, G.; Sun, X. Trigger is Non-central: Jointly event extraction via label-aware representations with multi-task learning. *Knowl.-Based Syst.* **2022**, *252*, 109480. [CrossRef]
110. Wadden, D.; Wennberg, U.; Luan, Y.; Hajishirzi, H. Entity, Relation, and Event Extraction with Contextualized Span Representations. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, Hong Kong, China, 3–7 November 2019; pp. 5783–5788.
111. Balali, A.; Asadpour, M.; Campos, R.; Jatowt, A. Joint event extraction along shortest dependency paths using graph convolutional networks. *Knowl.-Based Syst.* **2020**, *210*, 106492. [CrossRef]
112. Zhang, T.; Whitehead, S.; Zhang, H.; Li, H.; Ellis, J.G.; Huang, L.; Liu, W.; Ji, H.; Chang, S. Improving Event Extraction via Multimodal Integration. In Proceedings of the 2017 ACM on Multimedia Conference, Mountain View, CA, USA, 23–27 October 2017; pp. 270–278.
113. Li, Q.; Ji, H.; Huang, L. Joint Event Extraction via Structured Prediction with Global Features. In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, Sofia, Bulgaria, 4–9 August 2013; pp. 73–82.
114. Chen, B.; Lin, X.; Thomas, C.; Li, M.; Yoshida, S.; Chum, L.; Ji, H.; Chang, S. Joint Multimedia Event Extraction from Video and Article. In Proceedings of the Findings of the Association for Computational Linguistics, Online Event, 1–6 August 2021; pp. 74–88.
115. Sadhu, A.; Gupta, T.; Yatskar, M.; Nevatia, R.; Kembhavi, A. Visual Semantic Role Labeling for Video Understanding. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Virtual, 19–25 June 2021; pp. 5589–5600.
116. Chen, B.; Lin, X.; Thomas, C.; Li, M.; Yoshida, S.; Chum, L.; Ji, H.; Chang, S.F. Joint Multimedia Event Extraction from Video and Article. *arXiv* **2021**, arXiv:2109.12776.
117. Shen, W.; Wang, J.; Han, J. Entity Linking with a Knowledge Base: Issues, Techniques, and Solutions. *IEEE Trans. Knowl. Data Eng.* **2015**, *27*, 443–460. [CrossRef]
118. Sevgili, Ö.; Shelmanov, A.; Arkhipov, M.Y.; Panchenko, A.; Biemann, C. Neural Entity Linking: A Survey of Models based on Deep Learning. *arXiv* **2020**, arXiv:2006.00575.
119. Le, P.; Titov, I. Distant Learning for Entity Linking with Automatic Noise Detection. In Proceedings of the 57th Conference of the Association for Computational Linguistics, Florence, Italy, 28 July–2 August 2019; pp. 4081–4090.
120. Moreno, J.G.; Besançon, R.; Beaumont, R.; D'hondt, E.; Ligozat, A.; Rosset, S.; Tannier, X.; Grau, B. Combining Word and Entity Embeddings for Entity Linking. In *Semantic Web—14th International Conference, Portorož, Slovenia, 28 May–1 June 2017*; Springer: Cham, Switzerland, 2017; pp. 337–352.
121. Zwicklbauer, S.; Seifert, C.; Granitzer, M. Robust and Collective Entity Disambiguation through Semantic Embeddings. In Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval, Pisa, Italy, 17–21 July 2016; pp. 425–434.
122. Pershina, M.; He, Y.; Grishman, R. Personalized Page Rank for Named Entity Disambiguation. In Proceedings of the NAACL HLT 2015, The 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Denver, CO, USA, 4–5 June 2015; pp. 238–243.
123. Onoe, Y.; Durrett, G. Fine-Grained Entity Typing for Domain Independent Entity Linking. In Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, New York, NY, USA, 7–12 February 2020; pp. 8576–8583.

124. Shahbazi, H.; Fern, X.Z.; Ghaeini, R.; Ma, C.; Obeidat, R.; Tadepalli, P. Joint Neural Entity Disambiguation with Output Space Search. *arXiv* **2018**, arXiv:1806.07495.
125. Francis-Landau, M.; Durrett, G.; Klein, D. Capturing Semantic Similarity for Entity Linking with Convolutional Neural Networks. In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego, CA, USA, 12–17 June 2016; pp. 1256–1261.
126. Nguyen, T.H.; Fauceglia, N.R.; Muro, M.R.; Hassanzadeh, O.; Gliozzo, A.; Sadoghi, M. Joint learning of local and global features for entity linking via neural networks. In Proceedings of the COLING 2016, the 26th International Conference on Computational Linguistics, Osaka, Japan, 11–16 December 2016; pp. 2310–2320.
127. Ganea, O.E.; Hofmann, T. Deep Joint Entity Disambiguation with Local Neural Attention. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Copenhagen, Denmark, 7–11 September 2017; pp. 2619–2629.
128. Gupta, N.; Singh, S.; Roth, D. Entity Linking via Joint Encoding of Types, Descriptions, and Context. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Copenhagen, Denmark, 7–11 September 2017; pp. 2681–2690.
129. Eshel, Y.; Cohen, N.; Radinsky, K.; Markovitch, S.; Yamada, I.; Levy, O. Named Entity Disambiguation for Noisy Text. In Proceedings of the 21st Conference on Computational Natural Language Learning, Vancouver, BC, Canada, 3–4 August 2017.
130. Wu, L.; Petroni, F.; Josifoski, M.; Riedel, S.; Zettlemoyer, L. Scalable Zero-shot Entity Linking with Dense Entity Retrieval. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, Online, 16–20 November 2020; pp. 6397–6407.
131. Yamada, I.; Washio, K.; Shindo, H.; Matsumoto, Y. Global Entity Disambiguation with Pretrained Contextualized Embeddings of Words and Entities. *Globalization* **2021**. [CrossRef]
132. Huang, H.; Heck, L.P.; Ji, H. Leveraging Deep Neural Networks and Knowledge Graphs for Entity Disambiguation. *arXiv* **2015**, arXiv:1504.07678.
133. Cao, Y.; Huang, L.; Ji, H.; Chen, X.; Li, J. Bridge Text and Knowledge by Learning Multi-Prototype Entity Mention Embedding. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, Vancouver, BC, Canada, 30 July–4 August 2017; pp. 1623–1633.
134. Fang, W.; Zhang, J.; Wang, D.; Chen, Z.; Li, M. Entity Disambiguation by Knowledge and Text Jointly Embedding. In Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning, Berlin, Germany, 11–12 August 2016; pp. 260–269.
135. Radhakrishnan, P.; Talukdar, P.P.; Varma, V. ELDEN: Improved Entity Linking Using Densified Knowledge Graphs. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, New Orleans, LA, USA, 1–6 June 2018; pp. 1844–1853.
136. Perozzi, B.; Al-Rfou, R.; Skiena, S. DeepWalk: Online learning of social representations. In Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York, NY, USA, 24–27 August 2014; pp. 701–710.
137. Banerjee, D.; Chaudhuri, D.; Dubey, M.; Lehmann, J. PNEL: Pointer Network Based End-To-End Entity Linking over Knowledge Graphs. In *Semantic Web—ISWC 2020—19th International Semantic Web Conference, Athens, Greece, 2–6 November 2020*; Springer: Cham, Switzerland, 2020; pp. 21–38.
138. Nedelchev, R.; Chaudhuri, D.; Lehmann, J.; Fischer, A. End-to-End Entity Linking and Disambiguation leveraging Word and Knowledge Graph Embeddings. *arXiv* **2020**, arXiv:2002.11143.
139. Gillick, D.; Kulkarni, S.; Lansing, L.; Presta, A.; Baldridge, J.; Ie, E.; García-Olano, D. Learning Dense Representations for Entity Retrieval. In Proceedings of the 23rd Conference on Computational Natural Language Learning, Hong Kong, China, 3–4 November 2019; pp. 528–537.
140. Lazic, N.; Subramanya, A.; Ringgaard, M.; Pereira, F. Plato: A Selective Context Model for Entity Resolution. *Trans. Assoc. Comput. Linguist.* **2015**, *3*, 503–515. [CrossRef]
141. Peters, M.E.; Neumann, M.; IV, R.L.L.; Schwartz, R.; Joshi, V.; Singh, S.; Smith, N.A. Knowledge Enhanced Contextual Word Representations. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, Hong Kong, China, 3–7 November 2019; pp. 43–54.
142. Kolitsas, N.; Ganea, O.E.; Hofmann, T. End-to-End Neural Entity Linking. In Proceedings of the 22nd Conference on Computational Natural Language Learning, Brussels, Belgium, 31 October–1 November 2018; pp. 519–529.
143. Martins, P.H.; Marinho, Z.; Martins, A.F. Joint learning of named entity recognition and entity linking. *arXiv* **2019**, arXiv:1907.08243.
144. Moon, S.; Neves, L.; Carvalho, V. Multimodal Named Entity Disambiguation for Noisy Social Media Posts. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, Melbourne, Australia, 15–20 July 2018; pp. 2000–2008.
145. Adjali, O.; Besançon, R.; Ferret, O.; Le Borgne, H.; Grau, B. Building a Multimodal Entity Linking Dataset From Tweets. In Proceedings of the 12th Language Resources and Evaluation Conference, Marseille, France, 11–16 May 2020; pp. 4285–4292.
146. Gan, J.; Luo, J.; Wang, H.; Wang, S.; He, W.; Huang, Q. Multimodal Entity Linking: A New Dataset and A Baseline. In Proceedings of the 29th ACM International Conference on Multimedia, Chengdu, China, 20–24 October 2021; pp. 993–1001.
147. Wang, X.; Tian, J.; Gui, M.; Li, Z.; Wang, R.; Yan, M.; Chen, L.; Xiao, Y. WikiDiverse: A Multimodal Entity Linking Dataset with Diversified Contextual Topics and Entity Types. *arXiv* **2022**, arXiv:2204.06347.
148. Pagliardini, M.; Gupta, P.; Jaggi, M. Unsupervised Learning of Sentence Embeddings Using Compositional n-Gram Features. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, New Orleans, LA, USA, 1–6 June 2018; pp. 528–540.

149. Zheng, Q.; Wen, H.; Wang, M.; Qi, G. Visual Entity Linking via Multi-modal Learning. *Data Intell.* **2021**, *4*, 1–24. [CrossRef]
150. Wang, Q.; Mao, Z.; Wang, B.; Guo, L. Knowledge Graph Embedding: A Survey of Approaches and Applications. *IEEE Trans. Knowl. Data Eng.* **2017**, *29*, 2724–2743. [CrossRef]
151. Berant, J.; Chou, A.; Frostig, R.; Liang, P. Semantic Parsing on Freebase from Question-Answer Pairs. In Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, Seattle, WA, USA, 18–21 October 2013; pp. 1533–1544.
152. Weston, J.; Bordes, A.; Yakhnenko, O.; Usunier, N. Connecting Language and Knowledge Bases with Embedding Models for Relation Extraction. In Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, Seattle, WA, USA, 18–21 October 2013; pp. 1366–1371.
153. Riedel, S.; Yao, L.; McCallum, A.; Marlin, B.M. Relation Extraction with Matrix Factorization and Universal Schemas. In Proceedings of the Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics. The Association for Computational Linguistics, Atlanta, GA, USA, 9–14 June 2013; pp. 74–84.
154. Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G.S.; Dean, J. Distributed Representations of Words and Phrases and their Compositionality. In Proceedings of the Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems, Lake Tahoe, NV, USA, 5–8 December 2013; pp. 3111–3119.
155. Wang, Z.; Zhang, J.; Feng, J.; Chen, Z. Knowledge Graph Embedding by Translating on Hyperplanes. In Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence, Québec City, QC, Canada, 27–31 July 2014; pp. 1112–1119.
156. He, S.; Liu, K.; Ji, G.; Zhao, J. Learning to Represent Knowledge Graphs with Gaussian Embedding. In Proceedings of the 24th ACM International Conference on Information and Knowledge Management, Melbourne, Australia, 18–23 October 2015; pp. 623–632.
157. Nickel, M.; Tresp, V.; Kriegel, H. A Three-Way Model for Collective Learning on Multi-Relational Data. In Proceedings of the 28th International Conference on Machine Learning, Bellevue, WA, USA, 28 June–2 July 2011; pp. 809–816.
158. Yang, B.; Yih, W.; He, X.; Gao, J.; Deng, L. Embedding Entities and Relations for Learning and Inference in Knowledge Bases. In Proceedings of the 3rd International Conference on Learning Representations, San Diego, CA, USA, 7–9 May 2015.
159. Trouillon, T.; Welbl, J.; Riedel, S.; Gaussier, É.; Bouchard, G. Complex Embeddings for Simple Link Prediction. In Proceedings of the 33rd International Conference on Machine Learning, New York, NY, USA, 20–22 June 2016; pp. 2071–2080.
160. Dettmers, T.; Minervini, P.; Stenetorp, P.; Riedel, S. Convolutional 2D Knowledge Graph Embeddings. In Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018; pp. 1811–1818.
161. Nguyen, D.Q.; Nguyen, T.D.; Nguyen, D.Q.; Phung, D.Q. A Novel Embedding Model for Knowledge Base Completion Based on Convolutional Neural Network. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, New Orleans, LA, USA, 1–6 June 2018; pp. 327–333.
162. Yao, L.; Mao, C.; Luo, Y. KG-BERT: BERT for Knowledge Graph Completion. *arXiv* **2019**, arXiv:1909.03193.
163. Hinton, G.E.; Osindero, S.; Teh, Y.W. A Fast Learning Algorithm for Deep Belief Nets. *Neural Comput.* **2006**, *18*, 1527–1554. [CrossRef]
164. Salakhutdinov, R.; Hinton, G.E. Deep Boltzmann Machines. In Proceedings of the Twelth International Conference on Artificial Intelligence and Statistics, Clearwater Beach, FL, USA, 16–18 April 2009; Volume 5, pp. 1967–2006.
165. Srivastava, N.; Salakhutdinov, R. Learning representations for multimodal data with deep belief nets. In Proceedings of the International Conference on Machine Learning Workshop, Edinburgh, UK, 26 June–1 July 2012; Volume 79, p. 3.
166. Ngiam, J.; Khosla, A.; Kim, M.; Nam, J.; Lee, H.; Ng, A.Y. Multimodal deep learning. In Proceedings of the ICML, Bellevue, WA, USA, 28 June–2 July 2011.
167. Silberer, C.; Lapata, M. Learning grounded meaning representations with autoencoders. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, Baltimore, MD, USA, 23–24 June 2014; Volume 1, pp. 721–732.
168. Wang, D.; Cui, P.; Ou, M.; Zhu, W. Deep multimodal hashing with orthogonal regularization. In Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence, Buenos Aires, Argentina, 25–31 July 2015.
169. Feng, F.; Wang, X.; Li, R. Cross-modal retrieval with correspondence autoencoder. In Proceedings of the 22nd ACM International Conference on Multimedia, Lisboa, Portugal, 10–14 October 2014; pp. 7–16.
170. Wang, W.; Ooi, B.C.; Yang, X.; Zhang, D.; Zhuang, Y. Effective multi-modal retrieval based on stacked auto-encoders. *VLDB Endow.* **2014**, *7*, 649–660. [CrossRef]
171. Liu, Y.; Feng, X.; Zhou, Z. *Multimodal Video Classification with Stacked Contractive Autoencoders*; Elsevier: Amsterdam, The Netherlands, 2016; Volume 120, pp. 761–766.
172. Hong, C.; Yu, J.; Wan, J.; Tao, D.; Wang, M. Multimodal deep autoencoder for human pose recovery. *IEEE Trans. Image Process.* **2015**, *24*, 5659–5670. [CrossRef]
173. Hori, C.; Hori, T.; Lee, T.Y.; Zhang, Z.; Harsham, B.; Hershey, J.R.; Marks, T.K.; Sumi, K. Attention-based multimodal fusion for video description. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 4193–4202.

174. Lu, J.; Xiong, C.; Parikh, D.; Socher, R. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 375–383.
175. Chen, K.; Bui, T.; Fang, C.; Wang, Z.; Nevatia, R. AMC: Attention guided multi-modal correlation learning for image search. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2644–2652.
176. Long, X.; Gan, C.; Melo, G.; Liu, X.; Li, Y.; Li, F.; Wen, S. Multimodal keyless attention fusion for video classification. In Proceedings of the AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018; Volume 32.
177. Zadeh, A.; Liang, P.P.; Mazumder, N.; Poria, S.; Cambria, E.; Morency, L.P. Memory fusion network for multi-view sequential learning. In Proceedings of the AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018; Volume 32.
178. Lu, J.; Yang, J.; Batra, D.; Parikh, D. Hierarchical question-image co-attention for visual question answering. In Proceedings of the Advances in Neural Information Processing Systems 2016, Barcelona, Spain, 5–10 December 2016; Volume 29.
179. Zadeh, A.; Liang, P.P.; Poria, S.; Vij, P.; Cambria, E.; Morency, L.P. Multi-attention recurrent network for human communication comprehension. In Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018.
180. Reed, S.; Akata, Z.; Yan, X.; Logeswaran, L.; Schiele, B.; Lee, H. Generative adversarial text to image synthesis. In Proceedings of the International Conference on Machine Learning, New York, NY, USA, 19–24 June 2016; pp. 1060–1069.
181. Zhang, H.; Xu, T.; Li, H.; Zhang, S.; Wang, X.; Huang, X.; Metaxas, D.N. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 5907–5915.
182. Reed, S.E.; Akata, Z.; Mohan, S.; Tenka, S.; Schiele, B.; Lee, H. Learning what and where to draw. *arXiv* **2016**, arXiv:1610.02454.
183. Peng, Y.; Qi, J. *CM-GANs: Cross-Modal Generative Adversarial Networks for Common Representation Learning*; ACM: New York, NY, USA, 2019; Volume 15, pp. 1–24.
184. Xu, X.; He, L.; Lu, H.; Gao, L.; Ji, Y. *Deep Adversarial Metric Learning for Cross-Modal Retrieval*; Springer: Berlin/Heidelberg, Germany, 2019; Volume 22, pp. 657–672.
185. Zhang, J.; Peng, Y.; Yuan, M. Unsupervised generative adversarial cross-modal hashing. In Proceedings of the AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018; Volume 32.
186. Zhu, J.Y.; Park, T.; Isola, P.; Efros, A.A. Unpaired image-to-image translation using cycle-consistent adversarial networks. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2223–2232.
187. Wu, L.; Wang, Y.; Shao, L. Cycle-consistent deep generative hashing for cross-modal retrieval. *arXiv* **2018**, arXiv:1804.11013.
188. Sergieh, H.M.; Botschen, T.; Gurevych, I.; Roth, S. A Multimodal Translation-Based Approach for Knowledge Graph Representation Learning. In Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics, New Orleans, LA, USA, 5–6 June 2018; pp. 225–234.
189. Wang, Z.; Li, L.; Li, Q.; Zeng, D. Multimodal Data Enhanced Representation Learning for Knowledge Graphs. In Proceedings of the International Joint Conference on Neural Networks, Budapest, Hungary, 14–19 July 2019; pp. 1–8.
190. Pezeshkpour, P.; Chen, L.; Singh, S. Embedding Multimodal Relational Data for Knowledge Base Completion. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, 31 October–4 November 2018; pp. 3208–3218.
191. Pingali, S.; Yadav, S.; Dutta, P.; Saha, S. Multimodal Graph-based Transformer Framework for Biomedical Relation Extraction. *arXiv* **2021**, arXiv:2107.00596.
192. Zhang, J.; Zhang, H.; Xia, C.; Sun, L. Graph-bert: Only attention is needed for learning graph representations. *arXiv* **2020**, arXiv:2001.05140.
193. Zhang, H.; Fang, Q.; Qian, S.; Xu, C. Multi-Modal Knowledge-Aware Event Memory Network for Social Media Rumor Detection. In Proceedings of the 27th ACM International Conference on Multimedia, Nice, France, 21–25 October 2019; pp. 1942–1951. [CrossRef]
194. Wang, Y.; Qian, S.; Hu, J.; Fang, Q.; Xu, C. Fake News Detection via Knowledge-Driven Multimodal Graph Convolutional Networks. In Proceedings of the 2020 International Conference on Multimedia Retrieval, Dublin, Ireland, 8–11 June 2020; pp. 540–547.