Multi-Hop Open-Domain Question Answering over Structured and Unstructured Knowledge

Yue Feng, Zhen Han, Mingming Sun, Ping Li
Cognitive Computing Lab
Baidu Research
No. 10 Xibeiwang East Road, Beijing 10193, China
10900 NE 8th St, Bellevue, Washington 98004, USA
{fengyue, hanzhen, sunmingming01, liping11}@baidu.com

Abstract

Open-domain question answering systems need to answer question of our interests with structured and unstructured information. However, existing approaches only select one source to generate answer or only conduct reasoning on structured information. In this paper, we propose a Document-Entity Heterogeneous Graph Network, referred to as DEHG, to effectively integrate different sources of information, and conduct reasoning on heterogeneous information. DEHG employs a graph constructor to integrate structured and unstructured information, a context encoder to represent nodes and question, a heterogeneous information reasoning layer to conduct multi-hop reasoning on both information sources, and an answer decoder to generate answers for the question. Experimental results on HybirdQA dataset show that DEHG outperforms the state-of-the-art methods.

1 Introduction

Open-domain question answering (ODQA) is a task to answer any form of question in general domains with provided evidence (Chen and Yih, 2020; Sun et al., 2019, 2018b). The evidence that is used can be categorized into unstructured text like Wikipedia passages (Yang et al., 2018; Min et al., 2020; Izacard and Grave, 2021) and structured data like WikiData/WikiTables (Pasupat and Liang, 2015; Chen et al., 2020b; Wang et al., 2020; Feng et al., 2022). In practice, an ideal ODQA model should be able to analyze evidence from both unstructured text and structured data sources, as both types of evidence have their own advantages: 1) the unstructured text covers more general domains; 2) the structured data has better explainability to solve complex multi-hop reasoning.

One line of research accesses unstructured text and structured data independently (Sun et al., 2019; Xiong et al., 2019; Pan et al., 2021; Eisenschlos et al., 2021). The input question is sent to unstructured text system (TextQA) and structured knowledge base system (KBQA), and one of them is selected to output the final answer. These methods cannot combine the two sources of information properly. Recently, a new line of research aggregates heterogeneous information to find the answer (Chen et al., 2020b), which can construct connection between passages and table data. However, the method only conducts multi-hop reasoning on table data. It is difficult to handle questions that need to be answered when multi-hop reasoning on both sources is required.

In this work, we propose a novel Document-Entity Heterogeneous Graph Network (referred to as DEHG) for open-domain question answering which can conduct multi-hop reasoning on aggregated heterogeneous information. DEHG comprises a graph constructor to integrate heterogeneous information sources, a context encoder to generate representations for nodes and question, a heterogeneous information reasoning layer to explore multi-hoop connectivity of both information sources, and an answer decoder to generate answers for the question.

Our contributions can be summarized as follows: (1) we examine how to homogenize structured and unstructured knowledge in open-domain question answering for multi-hop reasoning. To the best of our knowledge, our work is the first to conduct multi-hop reasoning on integrated heterogeneous information in open-domain question answering. (2) We propose a Document-Entity Heterogeneous Graph Network to analyze complex relation of heterogeneous information in open-domain question answering. (3) We present experimental results that show DEHG outperforms previous state-of-the-art on HybirdQA dataset. We also perform an ablation study of our model to provide further insights.

Findings of the Association for Computational Linguistics: NAACL 2022, pages 151 - 156

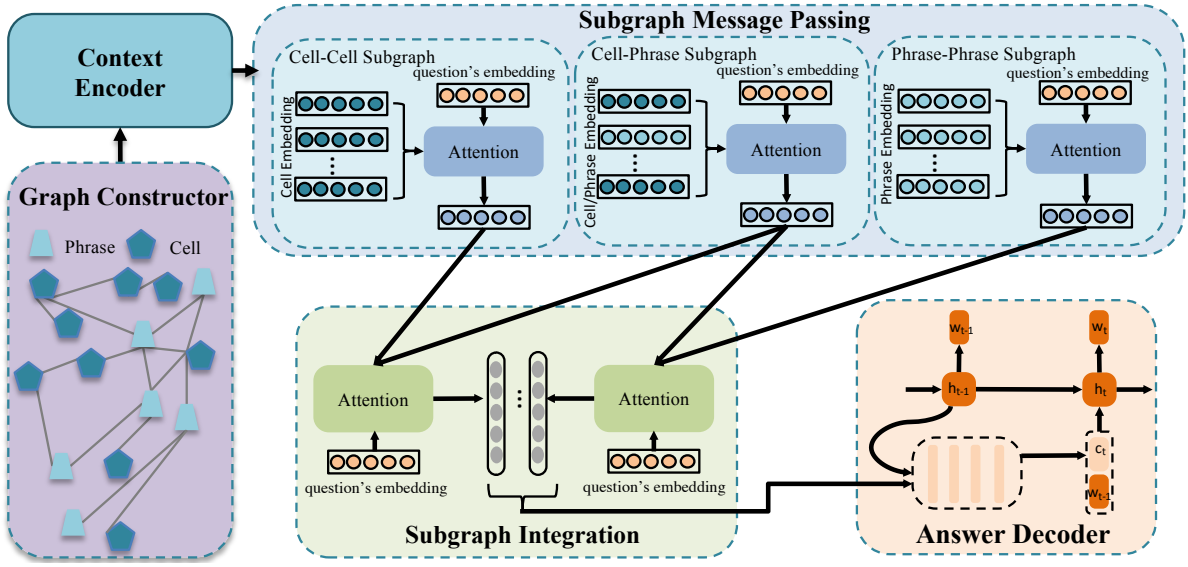

Figure 1: Overview of DEHG.

2 Our Approach

2.1 Graph Constructor

In order to cope with heterogeneous information, we propose a Document-Entity Heterogeneous Graph Constructor to enable rich heterogeneous information interaction. We divide the graph building process into two phases and describe them separately below:

Linking: This phase is aimed to link questions to their related information in tables and passages from two sources: 1) *Table Cell Matching*: in order to link related table cells to the question, we follow these three criteria: the table cell's value is explicitly mentioned by the question; the table cell's value is greater/less than the mentioned value in question; the table cell's value is maximum/minimum over the whole column if the question involves superlative words. 2) *Passage Matching*: it aims to link cells implicitly mentioned by the question through its hyperlinked passage. The linking model is a TF-IDF retriever with 3-gram lexicon which calculates the distances with all the passages in the pool and highlight the ones with distance lower than a threshold.

Building: this phase is aimed to build a heterogeneous graph to connect all linked cells and their corresponding hyperlinked passages. The structure of a heterogeneous graph is shown in Figure 1. For a heterogeneous graph $G = (V, E)$, V and E denote the set of nodes and the set of edges in the graph. The nodes V consist of the set of cells V_C, and the set of phrases of hyperlinked passages V_P. The edges E have three types, Cell-Cell edges E_{cc} that reflect the relations between cells, Cell-Phrase edges E_{cp} that describe the hyperlinked relation between cell and phrase, and Phrase-Phrase edges E_{pp} that express the semantic relation between phrases in the passage.

We utilize Open Information Annotation (OIA) (Sun et al., 2020), which is a predicate-function-argument annotation system for texts, to split passage into phrases and obtain the relation between phrases. Cells are connected to root phrase of its corresponding hyperlinked passage. All selected cells are connected to transfer information between cells on the heterogeneous graph.

2.2 Context Encoder

We use a BERT encoder to generate representations for every table cell, phrase of passage, and question as the initial node embedding in DEHG.

Each linked cell is encoded by 4-element tuple (CONTENT, LOCATION, SOURCE, SCORE). CONTENT represents the string representation in the table; LOCATION refers to the absolute row and column index in the table; SOURCE denotes where the entry comes from (e.g. equal/greater/less/min/max/passage); SCORE denotes the score of linked score normalized to [0, 1]. The first input token is [CLS], followed by the tokens of 4-element tuple, separated by [SEP]. The state of the first [CLS] is used as the cell's embedding s_c.

Each phrase in the passage is encoded by 2-

element tuple (TYPE, CONTENT). TYPE refers to the type of phrase extract by OIA (e.g. constant/predicate/function); CONTENT represents the sub-string in the passage; The input sequence starts with [CLS], followed by the tokens of 2-element tuple with [SEP] as a separator. The representation of [CLS] is used as the phrase's embedding s_p.

To generate the question's semantic embedding s_q, a BERT encoder is given the token sequence $X = ([\text{CLS}], x_1, ..., x_N, [\text{SEP}])$, where the sub-word tokens of the question are denoted as $x_1, ..., x_N$. [CLS] and [SEP] are start-of-text and separator pseudo-tokens respectively. The state of the first [CLS] is used as the question's embedding.

2.3 Heterogeneous Information Reasoning

Message passing: we define how information propagates over the graph in order to do reasoning over DEHG. According to the types of edges, the heterogeneous graph can be divided into three subgraphs: Cell-Cell subgraph, Cell-Phrase subgraph, and Phrase-Phrase subgraph. In each subgraph, we follow the message passing design in GCN (Kipf and Welling, 2017) to discriminate the importance of neighbors. To fuse the information of all subgraphs, we use the question-based attention to learn the corresponding weight of different subgraphs. With the learned weights as coefficients, we can fuse these subgraph embeddings to produce the finial node embedding.

Information Propagation: To explore the higher-order connectivity information of cells and passages, we stack T layers of subgraph representation and subgraph integration. Each layer k takes the node embedding from the previous layers as input, and outputs the updated node embedding after the current diffusion process finishes. The updated node embeddings are sent to the $k+1$ layer for the next diffusion process.

2.4 Answer Decoder

The state decoder sequentially generates the answer for the given question, which is represented as a sequence of pointers to cells of the tables and tokens of the passages. The pointers point to the nodes in the heterogeneous graph.

The state decoder is an LSTM using pointer (Vinyals et al., 2015) and attention (Bahdanau et al., 2015). It takes nodes semantic representations as input. At each decoding step t, the decoder receives the embedding of the previous item w_{t1}, the utterance context vector c_t, and the previous hidden state h_{t1}, and produces the current hidden state h_t,

$$h_t = \text{LSTM}(w_{t-1}, h_{t-1}, c_t). \quad (1)$$

We adopt the attention function in (Bahdanau et al., 2015)to calculate the context vectors as follows,

$$c_t = \text{atten}(h_{t-1}, N, N). \quad (2)$$

The decoder then generates a pointer from the set of pointers in the cells in the table and the phrases in the passages on the basis of the hidden state h_t. Specifically, it generates a pointer of item w according to the following distribution,

$$y_w = v^T \tanh(W_1 h_t + W_2 n_w), \quad (3)$$
$$P(w) = softmax(y_w), \quad (4)$$

where w is the pointer of node w, n_w is the representation of node w, v, W_1, and W_2 are trainable parameters, and softmax is calculated over all possible pointers.

3 Experiment

3.1 Dataset

We evaluate our multi-hop reasoning model DEHG on the HybridQA (Chen et al., 2020b) dataset, which contains factual questions that requires multi-hop reasoning using table and text. Tables and text are crawled from Wikipedia. Each row in the table describes several attributes of an instance. A table has its hyperlinked Wikipedia passages that describe the detail of attributes.

3.2 Baselines

In the following experiments, we compare our approach against previously published state-of-the-art approaches on the HybridQA dataset.

HyBrider (Chen et al., 2020b): A hybrid model that combines heterogeneous information to find the answer. *Unsupervised-QG* (Pan et al., 2021): An unsupervised framework that can generate questions by first selecting/generating relevant information from each data source. *DocHopper* (Sun et al., 2021): A multihop retrieval method that retrieves a paragraph or sentence. *Pointer* (Eisenschlos et al., 2021): A Transformer architecture that uses heads to attend to either rows or columns in a table.

3.3 Evaluation Measures

We use the following automatic evaluation metrics in our experiments. *Exact Match (EM)*: Measures what part of the predicted knowledge span matches the ground truth factoid exactly. *Token-Level F1*: We treat the predicted spans and ground truth factoids as bags of tokens, and compute F1.

3.4 Implementation Details

We use the pre-trained BERT model ([BERT-Base, Uncased]), which has 12 hidden layers of 768 units and 12 self-attention heads to encode cell, phrase, and question. The hidden size of LSTM decoder is also 768. The dropout probability is 0.1. We also use beam search for decoding, with a beam size of 5. The batch size is set to 4. Adam (Kingma and Ba, 2015) is used for optimization with an initial learning rate of 1e-4. We implement the algorithm using the PaddlePaddle Deep Learning Platform (Ma et al., 2019).

3.5 Experimental Results

In Table 1, we show the results of the our proposed DEHG graph based model on both development and test set and compare it with previously published results. It shows that our proposed DEHG works significantly better than the baselines in terms of EM and F1 on HybridQA. The results indicate that DEHG is really a general and effective model for multi-hop question answering over tabular and textual data. Specifically, DEHG can leverage the cell and phrase for question answering. It can also effectively handle multi-hop reasoning on the heterogeneous graph.

Model	Dev		Test	
	EM	F1	EM	F1
Unsupervised-QG	25.7	30.5	-	-
HyBrider	44.0	50.7	43.8	50.6
DocHopper	47.7	55.0	46.3	53.3
POINTR	63.4	71.0	62.8	70.2
DEHG	**65.2**	**76.3**	**63.9**	**75.5**

Table 1: Performance of our model and related work on the HybridQA dataset; Numbers in **bold** denote best results in that metric.

3.6 Ablation Study

We conduct ablation study on test set. We validate the effects of three factors: BERT-based encoder, heterogeneous information reasoning, and pointer generation decoder. The results indicate that all the components of DEHG are indispensable.

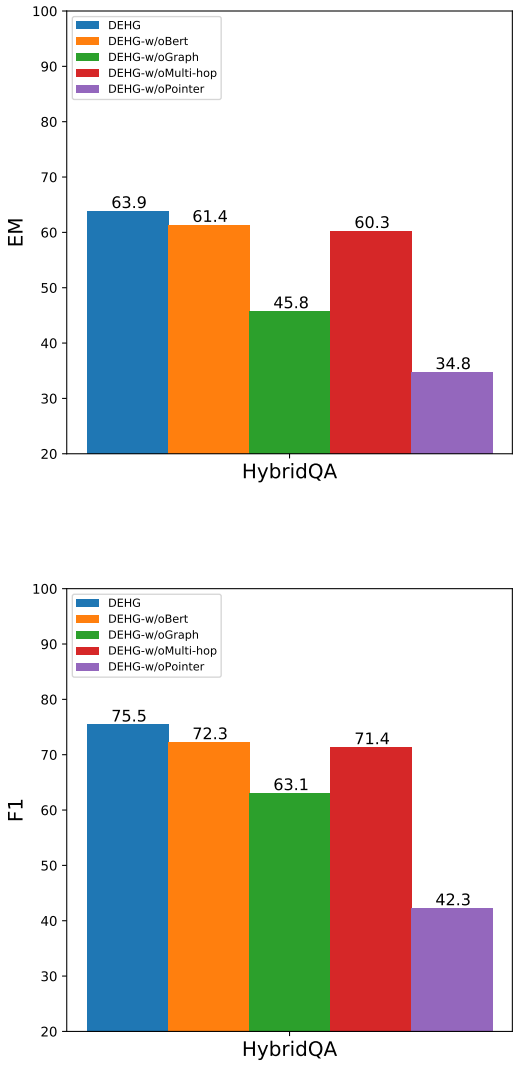

Figure 2: Ablation study results of DEHG.

Effect of BERT: To investigate the effectiveness of using BERT in the context encoder, we replace BERT with Bi-directional LSTM and run the model on HybridQA. As shown in Figure 2, the performance of the BiLSTM-based model DEHG-w/oBert in terms of EM and F1 decreases compared with DEHG. It indicates that the BERT-based context encoder can create and utilize more accurate representations for tabular and textual data and question understanding.

Effect of Heterogeneous Information Reasoning: To investigate the effectiveness of using the heterogeneous graph, we compare DEHG with DEHG-

w/oGraph which eliminates the heterogeneous information graph, and DEHG-w/oMulti-hop which removes the multi-hop information propagation. From Figure 2, one can observe that without the heterogeneous information graph the performances deteriorate considerably. In addition, the performances of DEHG-w/oGraph are inferior to DEHG-w/oMulti-hop. Thus, utilization of heterogeneous graph to representation multi-hop relation between passages and tables is desirable.

Effect of Pointer Decoder: To investigate the effectiveness of the pointer generation mechanism, we directly generate words from the vocabulary instead of generating pointers in the decoding process. Figure 2 also shows the results of DEHG-w/oPointer. From the results we can see that pointer generation is crucial for coping answer from cells and passages. It is due to HybridQA contains a large number of questions which answers are extracted from the tabular and textual data.

4 Related Work

Most work on QA uses structured and structured data independently (Talmor and Berant, 2018; Sun et al., 2018a; Kwiatkowski et al., 2019; Sun et al., 2019; Xiong et al., 2019; Chen et al., 2020a; Zhang et al., 2020; Liu et al., 2020; Pan et al., 2021; Eisenschlos et al., 2021; Yu et al., 2021). They use unstructured text system (TextQA) and structured knowledge base system (KBQA) to utilize different information. These methods cannot integrate different sources of information. A new method is proposed to aggregate heterogeneous information to find answer (Chen et al., 2020b; Feng et al., 2021). However, it only conducts multi-hop reasoning on table data. It is difficult to handle questions when multi-hop reasoning on both sources is required.

5 Conclusion

We have proposed a new approach to multi-hop question answering over tabular and textual data. The approach, referred to as DEHG, takes question answering as a problem of reasoning answers on the basis of a heterogeneous information graph. DEHG employs BERT in encoding of questions and passages respectively and generates pointers in decoding of answer generation. Experimental results show that DEHG significantly outperforms the state-of-the-art methods.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In Proceedings of the 3rd International Conference on Learning Representations (ICLR), San Diego, CA.

Danqi Chen and Wen-tau Yih. 2020. Open-domain question answering. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts (ACL), pages 34–37, Online.

Wenhu Chen, Hongmin Wang, Jianshu Chen, Yunkai Zhang, Hong Wang, Shiyang Li, Xiyou Zhou, and William Yang Wang. 2020a. Tabfact: A large-scale dataset for table-based fact verification. In International Conference on Learning Representations (ICLR).

Wenhu Chen, Hanwen Zha, Zhiyu Chen, Wenhan Xiong, Hong Wang, and William Yang Wang. 2020b. Hybridqa: A dataset of multi-hop question answering over tabular and textual data. In Findings of the Association for Computational Linguistics: (EMNLP), pages 1026–1036, Online Event.

Julian Eisenschlos, Maharshi Gor, Thomas Müller, and William W. Cohen. 2021. MATE: multi-view attention for table transformer efficiency. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 7606–7619, Virtual Event / Punta Cana, Dominican Republic.

Yue Feng, Aldo Lipani, Fanghua Ye, Qiang Zhang, and Emine Yilmaz. 2022. Dynamic schema graph fusion network for multi-domain dialogue state tracking. In Proceedings of the 60th Conference of the Association for Computational Linguistics (ACL).

Yue Feng, Zhaochun Ren, Weijie Zhao, Mingming Sun, and Ping Li. 2021. Multi-type textual reasoning for product-aware answer generation. In Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR), pages 1135–1145.

Gautier Izacard and Edouard Grave. 2021. Leveraging passage retrieval with generative models for open domain question answering. In Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume (EACL), pages 874–880, Online.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In Proceedings of the 3rd International Conference on Learning Representations (ICLR), San Diego, CA.

Thomas N. Kipf and Max Welling. 2017. Semi-supervised classification with graph convolutional networks. In Proceedings of the 5th International Conference on Learning Representations (ICLR), Toulon, France.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. 2019. Natural questions: a benchmark for question answering research. In Transactions of the Association for Computational Linguistics, volume 7, pages 453–466. MIT Press.

Guiliang Liu, Xu Li, Jiakang Wang, Mingming Sun, and Ping Li. 2020. Extracting knowledge from web text with monte carlo tree search. In Proceedings of the Web Conference (WWW), pages 2585–2591, Taipei.

Yanjun Ma, Dianhai Yu, Tian Wu, and Haifeng Wang. 2019. Paddlepaddle: An open-source deep learning platform from industrial practice. In Frontiers of Data and Domputing, volume 1, pages 105–115.

Sewon Min, Julian Michael, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2020. AmbigQA: Answering ambiguous open-domain questions. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 5783–5797, Online.

Liangming Pan, Wenhu Chen, Wenhan Xiong, Min-Yen Kan, and William Yang Wang. 2021. Unsupervised multi-hop question answering by question generation. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT), pages 5866–5880, Online.

Panupong Pasupat and Percy Liang. 2015. Compositional semantic parsing on semi-structured tables. 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 (ACL), pages 1470–1480, Beijing, China.

Haitian Sun, Tania Bedrax-Weiss, and William Cohen. 2019. PullNet: Open domain question answering with iterative retrieval on knowledge bases and text. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 2380–2390, Hong Kong, China.

Haitian Sun, William W Cohen, and Ruslan Salakhutdinov. 2021. End-to-end multihop retrieval for compositional question answering over long documents. arXiv preprint arXiv:2106.00200.

Mingming Sun, Wenyue Hua, Zoey Liu, Xin Wang, Kangjie Zheng, and Ping Li. 2020. A predicate-function-argument annotation of natural language for open-domain information eXpression. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 2140–2150, Online.

Mingming Sun, Xu Li, and Ping Li. 2018a. Logician and orator: Learning from the duality between language and knowledge in open domain. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 2119–2130, Brussels, Belgium.

Mingming Sun, Xu Li, Xin Wang, Miao Fan, Yue Feng, and Ping Li. 2018b. Logician: A unified end-to-end neural approach for open-domain information extraction. In Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining (WSDM), pages 556–564, Marina Del Rey, CA.

Alon Talmor and Jonathan Berant. 2018. The web as a knowledge-base for answering complex questions. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (NAACL), pages 641–651.

Oriol Vinyals, Meire Fortunato, and Navdeep Jaitly. 2015. Pointer networks. In Advances in Neural Information Processing Systems (NIPS), pages 2692–2700, Montreal, Canada.

Bailin Wang, Richard Shin, Xiaodong Liu, Oleksandr Polozov, and Matthew Richardson. 2020. RAT-SQL: relation-aware schema encoding and linking for text-to-sql parsers. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL), pages 7567–7578, Online.

Wenhan Xiong, Mo Yu, Shiyu Chang, Xiaoxiao Guo, and William Yang Wang. 2019. Improving question answering over incomplete kbs with knowledge-aware reader. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL), pages 4258–4264.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 2369–2380, Brussels, Belgium.

Jinxing Yu, Yunfeng Cai, Mingming Sun, and Ping Li. 2021. Mquade: a unified model for knowledge fact embedding. In Proceedings of the Web Conference (WWW), pages 3442–3452, Virtual Event / Ljubljana, Slovenia.

Jingyuan Zhang, Mingming Sun, Yue Feng, and Ping Li. 2020. Learning interpretable relationships between entities, relations and concepts via bayesian structure learning on open domain facts. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL), pages 8045–8056, Online.