Dynamic Semantic Graph Construction and Reasoning for Explainable Multi-hop Science Question Answering*

Weiwen Xu, Huihui Zhang, Deng Cai and Wai Lam
The Chinese University of Hong Kong
{wwxu,hhzhang,wlam}@se.cuhk.edu.hk
thisisjcykcd@gmail.com

Abstract

Knowledge retrieval and reasoning are two key stages in multi-hop question answering (QA) at web scale. Existing approaches suffer from low confidence when retrieving evidence facts to fill the knowledge gap and lack transparent reasoning process. In this paper, we propose a new framework to exploit more valid facts while obtaining explainability for multi-hop QA by dynamically constructing a semantic graph and reasoning over it. We employ Abstract Meaning Representation (AMR) as semantic graph representation. Our framework contains three new ideas: (a) AMR-SG, an AMR-based Semantic Graph, constructed by candidate fact AMRs to uncover any hop relations among question, answer and multiple facts. (b) A novel path-based fact analytics approach exploiting AMR-SG to extract active facts from a large fact pool to answer questions. (c) A fact-level relation modeling leveraging graph convolution network (GCN) to guide the reasoning process. Results on two scientific multi-hop QA datasets show that we can surpass recent approaches including those using additional knowledge graphs while maintaining high explainability on OpenBookQA and achieve a new state-of-the-art result on ARC-Challenge in a computationally practicable setting.

1 Introduction

Multi-hop QA is one of the most challenging tasks that benefits from explainability as it mimics the human question answering setting, where multi-hop QA requires both the collection of information from large external knowledge resources and the aggregation of retrieved facts to answer complex natural language questions (Yang et al., 2018).

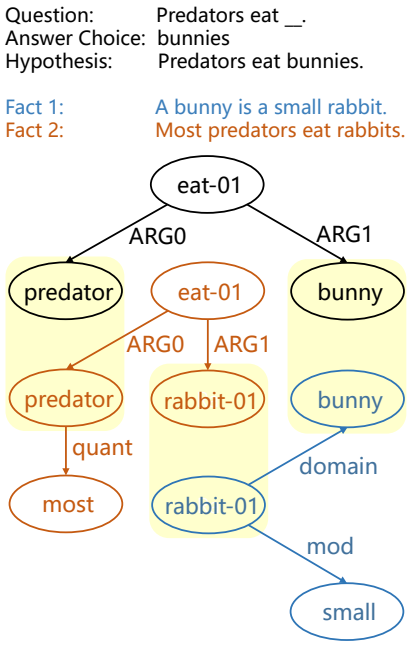

Figure 1: The AMR of the hypothesis (black), Fact 1 and Fact 2. A hypothesis is a statement derived from a question and a choice. The hypothesis AMR can be inferred by relevant fact AMRs.

Currently, external knowledge is mostly stored in two forms – textual and graph structure (e.g. Knowledge Graph (KG)). Textual corpora contain rich and diverse evidence facts, which are ideal knowledge resources for multi-hop QA. Especially with the success of pretrained models (Devlin et al., 2019; Liu et al., 2019; Lan et al., 2019), we can get powerful representations for such textual facts. However, retrieving relevant and useful facts to fill the knowledge gap for inferring the answer is still a challenging problem. In addition, the reasoning process over the facts is hidden by the unexplainable neural network, which hinders the deployment of real-life applications. On the other hand, KG is able to provide structural clues about relevant entities for explainable predictions (Feng et al.,

* The work described in this paper is substantially supported by a grant from the Research Grant Council of the Hong Kong Special Administrative Region, China (Project Code: 14204418).

Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pages 1044–1056
August 1–6, 2021.

2020; Saxena et al., 2020; Xu et al., 2020). But it is known to suffer from sparsity, where complex question clues are unlikely to be covered by the closed-form relations in KG (Zhao et al., 2020; Zhang et al., 2020b). Another issue is that KG requires large human labor and is easy to become outdated if not maintained timely.

To take advantages of both rich textual corpora and explicit graph structure and make it compatible to all textual knowledge, we explore the usefulness of Abstract Meaning Representation (AMR) as a graph annotation to a textual fact. AMR (Banarescu et al., 2013) is a semantic formalism that represents the meaning of a sentence into a rooted, directed graph. Figure 1 shows some examples of AMR graphs, where nodes represent concepts and edges represent the relations. Unlike other semantic role labeling that only considers the relations between predicates and their arguments (Song et al., 2019), the aim of AMR is to capture every meaningful content in high-level abstraction while removing away inflections and function words in a sentence. As a result, AMR allows us to explore textual facts and simultaneously attributes them with explicit graph structure for explainable fact quality assessment and reasoning.

In this paper, we propose a novel framework that incorporates AMR to make explainable knowledge retrieval and reasoning for multi-hop QA. Our framework works on textual knowledge, which is easy to obtain and allows us to get informative facts. The introduced AMR serves as a bridge that enables an explicit reasoning process over a graph structure among questions, answers and relevant facts. As exemplified in Figure 1, a hypothesis is first derived from a question and an answer choice. We then parse the hypothesis and a large number of facts to corresponding AMRs. After that, we dynamically construct AMR-SG for each question-choice pair by merging the AMRs of its hypothesis and relevant facts. Unlike previous works on multi-hop QA that rely on existing KGs to find relations among entities (Wang et al., 2020; Feng et al., 2020), our proposed AMR-SG is dynamically constructed, which reveals intrinsic relations of facts and can naturally form any-hop connections. After construction, we analyze all connected paths starting from the question to the answer on AMR-SG. We focus the consideration of facts on those paths because they together connect the question with the answer, indicating their active roles in filling the knowledge gap. The connections of facts on AMR-SG can be further used as the supervision for downstream reasoning. Therefore, we adopt GCN (Kipf and Welling, 2017) to model the fact-level information passing.

Experimental results demonstrate that our approach outperforms previous approaches that use additional KGs. It obtains 81.6 accuracy on OpenBookQA (Mihaylov et al., 2018), and pushes the state-of-the-art result on ARC-Challenge (Clark et al., 2018) to 68.94 in a computationally practicable setting.

2 Related Work

Multi-hop QA with External Resource. Despite the success of pretrained model in most Natural Language Processing (NLP) tasks, it performs poorly in multi-hop QA, where some information is missing to answer questions (Zhu et al., 2021b).

Textual corpora contain rich and diverse knowledge, which is likely to cover the clues to answer complex questions. Banerjee et al. (2019) demonstrate some carefully designed queries can effectively retrieve relevant facts. Yadav et al. (2019); Deng et al. (2020) extract groups of evidence facts considering the relevance, overlap and coverage, but such method requires exponential computation in the retrieval step. Feldman and El-Yaniv (2019); Yadav et al. (2020) construct a fact chain by iteratively reformulating the query to focus on the missing information. However, the fact chain often grows obliquely as a result of the failure of first fact retrieval, making the QA model brittle. As some recent QA datasets (Yang et al., 2018; Mihaylov et al., 2018; Khot et al., 2020) annotate a gold evidence fact for each question, it enables training supervised classifier to identify the correct fact driven by a query (Nie et al., 2019; Qiu et al., 2019; Tu et al., 2020; Banerjee and Baral, 2020). Min et al. (2018) take a further step to jointly predict the answer span and select evidence facts in a unified model. Though these supervised retrievers have achieved impressive improvement, they heavily rely on the annotated gold facts, which are not always available in real-world applications.

In addition, previous works also explore the effectiveness of structured knowledge by either encoding the nodes (Yang and Mitchell, 2017; Wang et al., 2019), triples (Mihaylov and Frank, 2018; Wang et al., 2020), paths (Lin et al., 2019; Lei et al., 2020) or tabular (Zhu et al., 2021a) to capture the

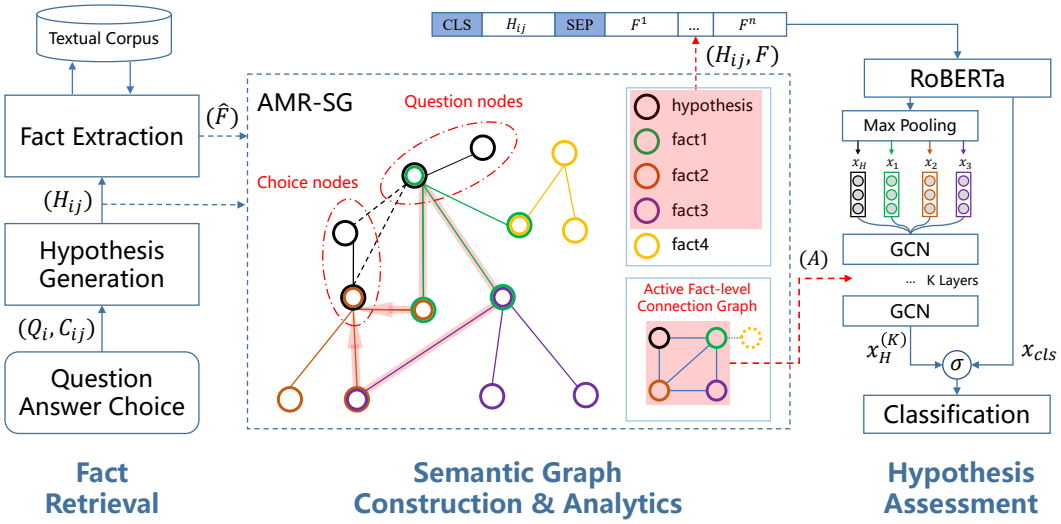

Figure 2: Overall architecture of our proposed model. The black dash lines in AMR-SG indicate that we cut the connection between question nodes and choice nodes. The pink arrows indicate two paths that can be spotted in AMR-SG. Facts with red background are active facts detected. The dashed node *Active Fact-level Connection Graph* indicates fact4 is not considered as a valid node as it is not an active fact.

missing information. Other works avoid the sparsity of KGs by constructing KGs directly from textual knowledge. OpenIE (Saha and Mausam, 2018) is widely used in knowledge base question answering to extract entity-relation triples (Bosselut et al., 2019; Zhao et al., 2020; Deng et al., 2019). However, OpenIE favors precision over recall, which is not necessarily effective to form connections among diverse evidence facts for multi-hop QA. Wikipedia contains internal hyperlinks, which are effective to build graph connections from unstructured articles (Asai et al., 2020; Liu et al., 2020). However, such hyperlinks are not available in most textual corpora.

AMR. Recent success in AMR research makes it possible to benefit downstream tasks, such as summarization (Takase et al., 2016; Dohare et al., 2017; Liao et al., 2018), event detection (Li et al., 2015) and machine translation (Song et al., 2019). In the domain of QA, AMR has been used to form logic queries and conduct symbolic reasoning (Mitra and Baral, 2016; Kapanipathi et al., 2020). Comparing to name entity (Zhong et al., 2020) or other cross-sentence annotations (Lei et al., 2018; Zhang et al., 2020a), we use AMR to build our semantic graph because it is align-free and can be easily adapted to powerful pretrained models.

3 Framework Description

In this paper, we consider the multi-hop QA in the form of multi-choice, where a question Q_i is provided with J answer choices $C_{ij}, j \in \{1, 2, ..., J\}$. As shown in Figure 2, our framework consists of three components: (1) a Fact Retrieval component to retrieve evidence facts $\hat{F} = \{\hat{F}^1, ..., \hat{F}^m\}$[1] for each question-choice pair from a large textual corpus; (2) a Semantic Graph Construction & Analytics component that dynamically constructs a semantic graph, named AMR-SG, to select active facts $F = \{F^1, ..., F^n\}$ from $\hat{F}$ and capture their relations A; and (3) a Hypothesis Assessment component that classifies whether the question-choice is correct, given the active facts and their relations in (2).

3.1 Fact Retrieval

Hypothesis Generation. As shown in Figure 2, we first generate a hypothesis H_{ij} for the i^{th} question and the j^{th} choice. A hypothesis is a completed statement derived from each question-choice pair. Comparing to simply concatenating the question and the choice, a hypothesis contains less meaningless words and maintain a good grammatical structure, which can avoid retrieving noisy facts and allow AMR parser to generate high-quality AMR graphs. We generate hypotheses by the rule-based model of Demszky et al. (2018). For some unsolvable cases, we directly concatenate the question and the choice. We apply this process for all training, develop and test sets.

[1] We omit the subscript ij for simplicity.

Fact Extraction. We retrieve a pool of evidence facts $\hat{F}$ for each hypothesis separately using *Elasticsearch* (Gormley and Tong, 2015). We set a large size m of the fact pool to cover as many valid facts as possible.

3.2 Semantic Graph Construction & Analytics

Active facts F are facts that really fill the knowledge gap between question and choice. The activeness of a fact cannot be simply determined by comparing it with the hypothesis, as multi-hop QA requires multiple facts to complete the reasoning chain. Therefore, we need to filter out facts that are just partially related and focus on the consideration of active facts and their roles in the reasoning chain. In this component, we first construct AMR-SG. Then, we propose a path-based analytics approach to extract active facts and construct an *Active Fact-level Connection Graph* to capture their relations with the question and the answer choice.

3.2.1 AMR-SG Construction

As the nodes of AMR are high-level abstraction of concepts conveyed in the corresponding textual fact, two AMRs sharing the same node indicate that they concern about the same concept, which shows their correlation. This motivates us to construct AMR-SG, shown in Figure 2, to represent the relations of the corresponding hypothesis and evidence facts for each question-choice pair.

We leverage the state-of-the-art AMR parser (Cai and Lam, 2020) to generate AMR $G = \{G^H, G^1, ..., G^m\}$ for a hypothesis and all facts in the corresponding fact pool, where G^H, G^i are the AMR of the hypothesis and the i^{th} fact respectively. AMR is also a directed and edge-labeled graph, which implies information specified in the edge is propagated in one predefined direction. However, such inner-AMR (edge labels and directions) information does not contribute to inter-AMR relations. Therefore, we only care about if there exists an edge between two nodes but ignore the edge labels and directions.

During construction, we regard G^H as the start point of AMR-SG. Then, we incrementally find one fact AMR in the fact pool sharing some nodes with it and add this fact AMR onto it by merging the shared nodes. The merging operation stops when no AMR can be added onto AMR-SG or the fact pool is empty. In fact, as shown in Figure 2, we do not change the architecture of each individual AMR, but reuse some shared nodes as the nodes in AMR-SG. Note that, some nodes are over-general, which are not appropriate to connect two AMRs (e.g. `(p/planet :name(n/name :op1"Earth"))`, the node `n/name` is an over-general concept). Fortunately, such over-general nodes always have non-node attributes (e.g. `Earth` of `n/name`) that shows the specific referent. Therefore, we replace the nodes with their non-node attributes if any to address this issue.

3.2.2 Path-based Analytics

Current multi-hop QA models are hindered by the quality of retrieved facts (Banerjee et al., 2019). We address this issue by a path-based analytics approach to guarantee the selected facts having a positive effect to answer the question.

As shown in Figure 2, AMR-SG reveals any-hop relations of the hypothesis and all facts. Completed paths can be spotted out of G^H to connect the question nodes with the choice nodes by passing through multiple facts. These facts, which together provide the missing knowledge to maintain complete reasoning chains, are active facts that we want to extract.

Specifically, we split the nodes of G^H into *question nodes* Q^H and *choice nodes* C^H. Question nodes represent the concepts extracted in the question text. As one question is provided with J choices, where we can generate J hypothesis AMRs. We take the shared nodes of these AMRs as Q^H, while the remaining as C^H:

$$Q^H_{ij} = \cap^J_{j=1}\{v|v \in G^H_{ij}\} \quad (1)$$

$$C^H_{ij} = \{v|v \in G^H_{ij}, v \notin Q^H_{ij}\}, j = 1, ..., J \quad (2)$$

We cut the edges between Q^H and C^H to guarantee the paths are spotted outside G^H. Then we apply depth-first search on AMR-SG to find all paths that connect at least one question node and one choice node, including the path that does not have a minimum length (e.g. the path passing through fact3 in Figure 2). All facts that the paths pass through (one node in and another node out) are considered as active facts. This is because we try to cover more facts as long as they do not deviate from the correct reasoning direction to provide enough information for QA model.

In addition, the any-hop relations of the hypothesis and active facts in AMR-SG can be used for a hypothesis to precisely aggregate knowledge from

relevant facts to reduce ambiguity during the reasoning process. Therefore, we construct an *Active Fact-level Connection Graph* from AMR-SG to capture such relations among the hypothesis and all active facts. As shown in Figure 2, each node in *Active Fact-level Connection Graph* is either the hypothesis or an active fact. We draw an edge between two facts (include hypothesis) if they share one concept node in AMR-SG.

3.3 Hypothesis Assessment with Fact-level Reasoning

As shown in Figure 2, we concatenate the hypothesis with all active facts, where [SEP] token is inserted between the two texts and [CLS] is put at the beginning of the sequence. We feed the whole sequence into a pretrained model based on RoBERTa (Liu et al., 2019) architecture to get the hidden representation of each token.

Then, *Active Fact-level Connection Graph* is used as an additional supervision in fact-level modeling to guide the reasoning process. Formally, let $s^H_{1:l_H} \in \mathbb{R}^{l_H \times d}$, $s^i_{1:l_i} \in \mathbb{R}^{l_i \times d}$ be the hidden representations of the hypothesis and the i^{th} active fact respectively, where l_H, l_i denote the length and d is the dimension of the representation. A max pooling layer is applied over these hidden representations to get the node representations respectively:

$$\begin{aligned} x_H &= \textbf{MaxPool}(s^H_{1:l_H}) \in \mathbb{R}^{1\times d} \\ x_i &= \textbf{MaxPool}(s^i_{1:l_i}) \in \mathbb{R}^{1\times d}, i = 1, ..., n \end{aligned} \tag{3}$$

The connections of hypothesis (0^{th}) and active facts in *Active Fact-level Connection Graph* can be viewed as an adjacency matrix $A \in \mathbb{R}^{(n+1)\times(n+1)}$, where

$$A_{ij} = \begin{cases} 1 & \text{if } F^i \text{ is connected with } F^j \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

As there is no edge information in the graph, a simple GCN is enough to model the knowledge fusion among the hypothesis and multiple active facts in the reasoning process. We also introduce multi-head mechanism (Vaswani et al., 2017) to stabilize the learning of different knowledge:

$$X^{(k)} = [head^{(k)}_1 : ... : head^{(k)}_h] \tag{5}$$

where [:] denotes concatenation operation, $X^{(k)}$ is the node states at the k^{th} layer, $X^{(0)} = [x_H; x_1; ...; x_n]$, [;] denotes the sequential concatenation operation, $head_i$ is the i^{th} head. Specifically, we compute the nodes states by aggregating knowledge from their neighboring nodes in each layer:

$$head^{(k)}_i = \text{ReLU}(\Lambda X^{(k-1)} W^{(k)}_i) \tag{6}$$

where $W^{(k)}_i \in \mathbb{R}^{d\times(d/h)}$ is the projection matrix of $head_i$ at the k^{th} layer, h is the head number. Λ is the normalization constant to avoid scale changing:

$$\begin{aligned} \Lambda &= D^{-1/2} A D^{-1/2} \\ D_{ii} &= \textstyle\sum_j A_{ij} \end{aligned} \tag{7}$$

After that, a σ gate is applied to calculate how much knowledge can be propagated to score the question-choice pair:

$$\lambda = \sigma(W^\lambda [x_{cls} : x^{(K)}_H] + b^\lambda) \tag{8}$$

$$s(q, a) = W^o(\lambda x^{(K)}_H + (1 - \lambda) x_{cls}) + b^o \tag{9}$$

where $W^\lambda \in \mathbb{R}^{1\times 2d}$, $W^o \in \mathbb{R}^{d\times d}$, b^λ, b^o are the parameters. We get the final probability by normalize all question-choice pairs with softmax.

	Train	Dev	Test
OpenBookQA	4957	500	500
ARC-Challenge	8992	299	1172

Table 1: Number of instances in each dataset.

4 Experiments

4.1 Datasets

We evaluate our approach on two multi-choice multi-hop QA datasets: ARC-Challenge (Clark et al., 2018) and OpenBookQA (Mihaylov et al., 2018). The textual corpus we use for both datsets is ARC Corpus (Clark et al., 2018), which contains about 14M science facts. OpenBookQA and ARC-Challenge have their leaderboards with train, develop and test sets publicly available. we follow AllenAI (2019) to combine the training set of OpenBookQA (4957), ARC-Easy (2251), ARC-Challenge (1119) and RegLivEnv (665) as the final training set of ARC-Challenge task. The data splits is shown in Table 1.

For ARC-Challenge, we retrieve 100 facts to form the fact pool. Based on this, we select up to 20 active facts using our approach as the context for each question-choice pair.[2] OpenBookQA provides an accompanying open-book of 1326 science

[2]We can only reproduce the results similar to AllenAI (2019) using 20 facts as the context.

facts, which are highly related to the questions in this dataset. Therefore, for OpenBookQA, we retrieve 10 facts from the open-book and another 90 facts from ARC Corpus, forming the 100 facts in the fact pool. We then select up to 15 active facts using our approach as the context.

4.2 Implementation

We implement our approach on two pretrained models: RoBERTa (Liu et al., 2019) and AristoRoBERTa (AllenAI, 2019). AristoRoBERTa employs the RoBERTa architecture but uses RACE (Lai et al., 2017) to first fine-tune the RoBERTa model. We prepare active facts as the context to further fine-tune the model with the target dataset. For OpenBookQA, we continue to fine-tune the QA model following the same procedure as AllenAI (2019), where the initial learning rate is 2e-5, the batch size is 12 and the max sequence length is 256. For ARC-Challenge, the initial learning rate, the batch size and the max sequence length are 1e-5, 6, and 416 respectively. We use grid search to find optimal hyper-parameters, where the learning rate is chosen from {5e-6, 1e-5, 2e-5}, the batch size is chosen from {4, 6, 8, 12, 16}. The number of GCN layer K is chosen from {1,**2**,3,4}, while the head number h is the RoBERTa-Large default value.[3]

We introduce 6M parameters of the fact-level reasoning module in addition to 355M of RoBERTa-Large. We run all experiments on one TITAN RTX card, which takes about 1 hour and 3 hours to complete the training of OpenBookQA and ARC-Challenge respectively.

4.3 Comparison Methods

We compare with recent existing methods that make use of similar power of pretrained models in order to conduct a fair comparison. These include the baseline AristoRoBERTaV7 (AllenAI, 2019) finetuned on top of AristoRoBERTa, KF-SIR (Banerjee and Baral, 2020) that exploits the knowledge fusion among facts, FreeLB (Zhu et al., 2020) that tackles the robustness issue and another three methods leveraging an additional knowledge graph (Speer et al., 2017) in addition to the textual knowledge: PG (Wang et al., 2020), MHGRN (Feng et al., 2020), AlBERT + KB. PG(albert + gpt2, roberta + gpt2) are two implementations with different pretrained model architectures (Liu et al., 2019; Lan et al., 2019; Radford et al., 2019), where the latter is more fair to compare with us.

[3]Our code is available at: https://github.com/wwxu21/AMR-SG

Methods	Model Architecture	Additional KG	Test Acc.
PG	albert + gpt2	✓	81.8
PG	roberta + gpt2	✓	80.2
AlBERT + KB	albert	✓	81.0
MHGRN	roberta	✓	80.6
KF-SIR	roberta	×	80.0
AristoRoBERTaV7	roberta	×	77.8
+ AMR-SG-Full	roberta	×	**81.6**

Table 2: Test accuracy on OpenBookQA. Methods using additional KG are ticked.

5 Results

5.1 Main Results

OpenBookQA. The test set accuracy is shown in Table 2. AMR-SG-Full is our full model based on AristoRoBERTa. Results show that AMR-SG-Full can surpass models leveraging additional KG. It demonstrates that the fundamental improvement of AMR-SG-Full comes from the knowledge mining of the textual corpus. However, such knowledge resource has not been fully investigated by existing methods and contains richer and more diverse evidence facts than KGs. We do not compare with UnifiedQA (Khashabi et al., 2020) and T5 3B (Raffel et al., 2020) as they rely on extremely large pretrained models (at least 3B parameters), which are not fair for comparison.

ARC-Challenge. We also implement AMR-SG-Full on another difficult multi-hop QA dataset: ARC-Challenge. It consists of the questions only answered incorrectly by both a retrieval-based algorithm and a word co-occurrence algorithm (Clark et al., 2018), which theoretically is not friendly to our approach. As shown in Table 3, we can still obtain 2.47 accuracy improvement comparing to AristoRoBERTaV7 and achieve a new state-of-the-art performance in a computationally practicable setting.

5.2 Ablation Study

We conduct ablation study by incrementally adding each component of AMR-SG-Full to investigate its effectiveness on two pretrained models in Table 4. We include the analysis on RoBERTa because it is a more general and widely used pretrained model.

Methods	Test Acc.
FreeLB (Zhu et al., 2020)	67.75
arcRoberta	67.15
xlnet+Roberta	67.06
AristoRoBERTaV7 (AllenAI, 2019)	66.47
+ AMR-SG-Full	**68.94**

Table 3: Test accuracy on ARC-Challenge. All models use RoBERTa architecture for the pretrained model and do not leverage additional KG.

Methods	Dev	Test
RoBERTa		
No Fact	66.8	64.8
+ Fact Context	68.2	68.8 (+4.0)
+ Fact Analytics	73.2	73.0 (+4.2)
+ Fact-level Reasoning	72.8	74.2 (+1.2)
AristoRoBERTa		
No Fact	71.0	70.0
+ Fact Context ♠	78.2	78.4 (+8.4)
+ Fact Analytics	79.4	81.4 (+3.0)
+ Fact-level Reasoning	**79.6**	**81.6** (+0.2)

Table 4: Ablation study of model components on OpenBookQA (adding one component incrementally). ♠ is our reimplementation of (AllenAI, 2019).

We start from the vanilla pretrained models, where no textual facts are provided (denoted as *No Fact*). We retrieve 15 facts as the context to create the first variant (denoted as + *Fact Context*). The purpose is to test the contribution of the facts retrieved by the simple information retrieval (IR) system (*Elasticsearch*). We continue to add the path-based fact analytics component (denoted as + *Fact Analytics*). In fact, this variant merely use the facts selected from `AMR-SG` to fine-tune the pretrained models. On top of both two pretrained models, we observe a great performance improvement, where the improvement brought by + *Fact Analytics* is higher than + *Fact Context* on top of RoBERTa, which demonstrates this component can effectively select useful facts to fill the knowledge gap that have not been covered by the IR system. We finally equip our model with the fact-level reasoning component (denoted as + *Fact-level Reasoning*). From the results, we can observe that this component performs well on top of RoBERTa, but has very little effect on top of AristoRoBERTa. This is because this component tries to infuse some fact-level connections to ease the reasoning process of the model. Such information can be learned automatically by the model itself if exposed to enough in-domain data (AristoRoBERTa). Nevertheless, the fact-level reasoning is a more general method when such data is unavailable.

6 Explainability Analysis

6.1 Analysis of AMR-SG

Impact of Evidence Facts. As discussed above, the major improvement of our approach comes from more useful facts selected for each question-choice pair. In this section, we take a deep look at the quality and the composition of those facts on OpenBookQA. We derive five variants by varying the composition of core (facts retrieved from open-book) or common (facts from ARC Corpus) facts. For core facts, as open-book annotates one gold core fact for each question, the retrieval accuracy of the gold fact is a natural way to evaluate the quality. For common facts, we conduct human analysis to evaluate the quality from three aspects: (1) Relatedness: Does the retrieved fact related to the question or the answer? (2) Informativeness: Does the retrieved fact provided useful information to answer the question? (3) Completeness: Do all retrieved facts together fill the knowledge gap to completely answer the question? We randomly sample 50 questions and evaluate the evidence facts corresponding to the correct answer choice, where one fact would contribute 1 score if it meets the requirement of Relatedness or Informativeness respectively and all 15 facts contribute 1 score if they together meet the requirement of Completeness. Evaluation results are presented in Table 5.

When varying the fact composition of IR variants, we find the gold core fact retrieval accuracy has a positive impact on the final accuracy on top of RoBERTa. At this stage, some questions can be inferred sufficiently with the gold core facts. Higher retrieval accuracy accounts for more questions of this kind to be correctly answered. However, this advantage is not as obvious for AristoRoBERTa. Our human evaluation reveals that such facts are unlikely to form a complete reasoning chain, making it hard for real multi-hop reasoning.

On the other hand, our approach directly models the intrinsic fact relations, where the path-based analytics ensures that the facts selected are in the reasoning chain from the question to the answer. Results show that our approach makes an overall improvement with regard to Relatedness, Informativeness and Completeness and is

Facts Composition (total 15 facts)	Core Fact Retrieval Accuracy	Human Evaluation			Test set Accuracy	
		Rel.	Info.	Comp.	RoBERTa	AristoRoBERTa
IR (5/10)	56.4	5.86	2.50	0.46	68.8	78.4
IR (10/5)	63.6	5.20	2.24	0.42	70.4	77.4
IR (15/0)	68.4	3.36	1.62	0.26	72.2	77.4
AMR-SG (10/30)	61.0	5.85	2.58	0.48	72.4	80.4
AMR-SG (10/100)	61.0	6.22	2.98	0.56	**74.2**	**81.6**

Table 5: Automatic and Human Evaluation of the evidence facts on OpenBookQA. IR (x/y) indicates we use simple IR system to retrieve x core facts and y common facts. AMR-SG (x/y) indicates we construct AMR-SG with x core facts and y common facts, based on which we then select 15 active facts and extract their relations.

less harmful to core fact retrieval. We also find that AMR-SG (10/100) can make a further improvement compared to AMR-SG (10/30) by including more facts to construct AMR-SG. It demonstrates that AMR-SG has the capability of detecting useful facts from a large and noisy fact pool, thus making up for the deficiency of the IR system.

Impact of AMR Consistency. We investigate the quality consistency of AMR graphs to see how it affects the construction of AMR-SG and thus affects the QA model. We prepare AMR in three consistency levels, where *Fully-Automatic* is generated by automatic AMR parser; *Mixed* is that we manually annotate the error-free AMRs for the core facts in open-book (1326 in total) and use the error-free core fact AMRs and other automatically generated AMRs to construct AMR-SG; *Error-Free-Adapted* is that we use the error-free AMRs annotated to fine-tune the AMR parser and use the tuned parser to generate AMR for all the remaining facts (including hypotheses and common facts, about 900k in total). The test set accuracy are 81.6, 80.2, 80.4 for *Fully-Automatic*, *Mixed* and *Error-Free-Adapted* respectively. It is interesting to note that using *Fully-Automatic* AMRs results in higher QA accuracy than *Mixed* and *Error-Free-Adapted*, where the latter two contain a mix of AMRs with different levels of quality. This phenomenon has also been observed in other AMR applications (Liu et al., 2015; Hardy and Vlachos, 2018), where automatic parses perform well than manual parses. We conjecture that this can be attributed to the discrepancy between the error-free AMRs and the automatically parsed AMRs in the choices of AMR concepts with similar meaning. This small difference in concept choices may omit potential connections, results in some important facts failing to be detected. In contrast, automatically parsed AMRs contain errors, but they are consistent in their concept choices, which is more likely for AMRs to form connections. The 0.2 accuracy improvement between *Mixed* and *Error-Free-Adapted* also demonstrates our assumption, since the parser is finetuned on the error-free AMRs, where its parsed AMRs should be more consistent with the error-free AMRs.

Question: *A seismograph can accurately describe* (A) how rough the footing will be (B) how bad the weather will be **(C) how stable the ground will be** (D) how shaky the horse will be
Useful facts retrieved by IR: N.A.
Additional facts from path-based analytics: A seismograph is a kind of tool for measuring the size of an earthquake. An earthquake is a shockwave travelling through the ground.
Relevant path in AMR-SG: seismograph→tool→measure-01→size-01→earthquake→ground

Table 6: A case study showing how our framework selects useful facts to completely fill the knowledge gap.

6.2 Case Study

Table 6 shows one case study of evidence facts selected by our framework. Since the important term *earthquake* is missing from the search query, the IR system assigns low retrieval scores for the two facts, causing a low ranking. However, the two facts can form a complete reasoning chain with the question and the answer via several concept nodes, where our approach can successfully extract the two facts despite the low retrieval scores. More cases can be found in Appendix A.1.

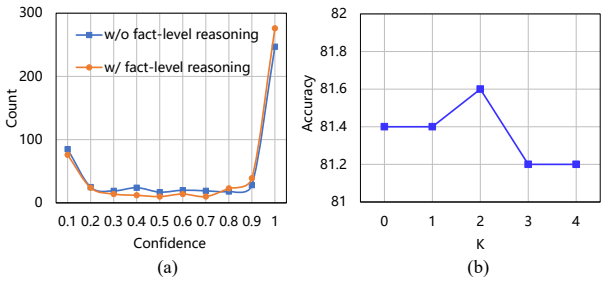

Figure 3: Analysis of fact-level reasoning on OpenBookQA. (a) presents the distribution of prediction confidence with or without fact-level reasoning module. (b) shows the QA performance with different GCN layer K. Size 0 denotes the original pretrained model.

6.3 Analysis of Fact-level Reasoning

Why Fact-level reasoning. Figure 3(a) shows that fact-level reasoning improves the performance by making a more confident prediction for the correct answer. This is because the fact-level connections of AMR−SG inform the model how these active facts are intrinsically related, which allows the model to precisely receive knowledge from related facts.

Impact of Number of Hops (K). We vary the hyper-parameter K to consider the impact of K-hop neighbors on OpenBookQA. As show in Figure 3(b), the performance reaches the top at $K = 2$. It indicates that most of the questions can be well answered using two evidence facts, which is consistent with the construction of this dataset. However, the performance drops when $K > 2$. It might be attributed to exponential noise found in longer reasoning chains.

7 Conclusion

We propose to dynamically construct AMR−SG that can reflect the intrinsic relations of relevant facts leveraging AMR, a graph annotation. AMR−SG combines the advantages of rich textual corpus and graph structure, where we can select useful facts that completely form the reasoning chain and make fact-level modeling. Experimental results show that AMR−SG can maintain high explainability, and successfully couple with strong pretrained models to achieve significant improvement on OpenBookQA and ARC-Challenge over approaches leveraging additional KGs.

References

AllenAI. 2019. Aristorobertav7. In *AristoRoBERTAv7*.

Akari Asai, Kazuma Hashimoto, Hannaneh Hajishirzi, Richard Socher, and Caiming Xiong. 2020. Learning to retrieve reasoning paths over wikipedia graph for question answering. In *International Conference on Learning Representations*.

Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. 2013. Abstract Meaning Representation for sembanking. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 178–186, Sofia, Bulgaria. Association for Computational Linguistics.

Pratyay Banerjee and Chitta Baral. 2020. Knowledge fusion and semantic knowledge ranking for open domain question answering. *arXiv preprint arXiv:2004.03101*.

Pratyay Banerjee, Kuntal Kumar Pal, Arindam Mitra, and Chitta Baral. 2019. Careful selection of knowledge to solve open book question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6120–6129, Florence, Italy. Association for Computational Linguistics.

Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi. 2019. COMET: Commonsense transformers for automatic knowledge graph construction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4762–4779, Florence, Italy. Association for Computational Linguistics.

Deng Cai and Wai Lam. 2020. AMR parsing via graph-sequence iterative inference. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1290–1301, Online. Association for Computational Linguistics.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*.

Dorottya Demszky, Kelvin Guu, and Percy Liang. 2018. Transforming question answering datasets into natural language inference datasets. *arXiv preprint arXiv:1809.02922*.

Yang Deng, Yuexiang Xie, Yaliang Li, Min Yang, Nan Du, Wei Fan, Kai Lei, and Ying Shen. 2019. Multi-task learning with multi-view attention for answer selection and knowledge base question answering. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019*, pages 6318–6325.

Yang Deng, Wenxuan Zhang, and Wai Lam. 2020. Multi-hop inference for question-driven summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020*, pages 6734–6744.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Shibhansh Dohare, Harish Karnick, and Vivek Gupta. 2017. Text summarization using abstract meaning representation. *arXiv preprint arXiv:1706.01678*.

Yair Feldman and Ran El-Yaniv. 2019. Multi-hop paragraph retrieval for open-domain question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2296–2309, Florence, Italy. Association for Computational Linguistics.

Yanlin Feng, Xinyue Chen, Bill Yuchen Lin, Peifeng Wang, Jun Yan, and Xiang Ren. 2020. Scalable multi-hop relational reasoning for knowledge-aware question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1295–1309, Online. Association for Computational Linguistics.

Clinton Gormley and Zachary Tong. 2015. *Elasticsearch: the definitive guide: a distributed real-time search and analytics engine*. " O'Reilly Media, Inc.".

Hardy Hardy and Andreas Vlachos. 2018. Guided neural language generation for abstractive summarization using Abstract Meaning Representation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 768–773, Brussels, Belgium. Association for Computational Linguistics.

Pavan Kapanipathi, Ibrahim Abdelaziz, Srinivas Ravishankar, Salim Roukos, Alexander Gray, Ramon Astudillo, Maria Chang, Cristina Cornelio, Saswati Dana, Achille Fokoue, et al. 2020. Question answering over knowledge bases by leveraging semantic parsing and neuro-symbolic reasoning. *arXiv preprint arXiv:2012.01707*.

Daniel Khashabi, Sewon Min, Tushar Khot, Ashish Sabharwal, Oyvind Tafjord, Peter Clark, and Hannaneh Hajishirzi. 2020. UNIFIEDQA: Crossing format boundaries with a single QA system. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1896–1907, Online. Association for Computational Linguistics.

Tushar Khot, Peter Clark, Michal Guerquin, Peter Jansen, and Ashish Sabharwal. 2020. Qasc: A dataset for question answering via sentence composition. In *AAAI*, pages 8082–8090.

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 '17.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017. RACE: Large-scale ReAding comprehension dataset from examinations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 785–794, Copenhagen, Denmark. Association for Computational Linguistics.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

Wenqiang Lei, Yuanxin Xiang, Yuwei Wang, Qian Zhong, Meichun Liu, and Min-Yen Kan. 2018. Linguistic properties matter for implicit discourse relation recognition: Combining semantic interaction, topic continuity and attribution. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Wenqiang Lei, Gangyi Zhang, Xiangnan He, Yisong Miao, Xiang Wang, Liang Chen, and Tat-Seng Chua. 2020. Interactive path reasoning on graph for conversational recommendation. *Proceedings of the 26th ACM SIGKDD*.

Xiang Li, Thien Huu Nguyen, Kai Cao, and Ralph Grishman. 2015. Improving event detection with Abstract Meaning Representation. In *Proceedings of the First Workshop on Computing News Storylines*, pages 11–15, Beijing, China. Association for Computational Linguistics.

Kexin Liao, Logan Lebanoff, and Fei Liu. 2018. Abstract Meaning Representation for multi-document summarization. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1178–1190, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Bill Yuchen Lin, Xinyue Chen, Jamin Chen, and Xiang Ren. 2019. KagNet: Knowledge-aware graph networks for commonsense reasoning. 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 2829–2839, Hong Kong, China. Association for Computational Linguistics.

Dayiheng Liu, Yeyun Gong, Jie Fu, Yu Yan, Jiusheng Chen, Daxin Jiang, Jiancheng Lv, and Nan Duan. 2020. Rikinet: Reading wikipedia pages for natural question answering.

Fei Liu, Jeffrey Flanigan, Sam Thomson, Norman Sadeh, and Noah A. Smith. 2015. Toward abstractive summarization using semantic representations. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1077–1086, Denver, Colorado. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. Can a suit of armor conduct electricity? a new dataset for open book question answering. In *EMNLP*.

Todor Mihaylov and Anette Frank. 2018. Knowledgeable reader: Enhancing cloze-style reading comprehension with external commonsense knowledge. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 821–832, Melbourne, Australia. Association for Computational Linguistics.

Sewon Min, Victor Zhong, Richard Socher, and Caiming Xiong. 2018. Efficient and robust question answering from minimal context over documents. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1725–1735, Melbourne, Australia. Association for Computational Linguistics.

Arindam Mitra and Chitta Baral. 2016. Addressing a question answering challenge by combining statistical methods with inductive rule learning and reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30.

Yixin Nie, Songhe Wang, and Mohit Bansal. 2019. Revealing the importance of semantic retrieval for machine reading at scale. 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 2553–2566, Hong Kong, China. Association for Computational Linguistics.

Lin Qiu, Yunxuan Xiao, Yanru Qu, Hao Zhou, Lei Li, Weinan Zhang, and Yong Yu. 2019. Dynamically fused graph network for multi-hop reasoning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6140–6150, Florence, Italy. Association for Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Swarnadeep Saha and Mausam. 2018. Open information extraction from conjunctive sentences. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2288–2299, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Apoorv Saxena, Aditay Tripathi, and Partha Talukdar. 2020. Improving multi-hop question answering over knowledge graphs using knowledge base embeddings. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4498–4507, Online. Association for Computational Linguistics.

Linfeng Song, Daniel Gildea, Yue Zhang, Zhiguo Wang, and Jinsong Su. 2019. Semantic neural machine translation using AMR. *Transactions of the Association for Computational Linguistics*, 7:19–31.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31.

Sho Takase, Jun Suzuki, Naoaki Okazaki, Tsutomu Hirao, and Masaaki Nagata. 2016. Neural headline generation on Abstract Meaning Representation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1054–1059, Austin, Texas. Association for Computational Linguistics.

Ming Tu, Kevin Huang, Guangtao Wang, Jing Huang, Xiaodong He, and Bowen Zhou. 2020. Select, answer and explain: Interpretable multi-hop reading comprehension over multiple documents. In *AAAI*, pages 9073–9080.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, pages 5998–6008. Curran Associates, Inc.

Peifeng Wang, Nanyun Peng, Filip Ilievski, Pedro Szekely, and Xiang Ren. 2020. Connecting the dots: A knowledgeable path generator for commonsense question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4129–4140, Online. Association for Computational Linguistics.

Xiaoyan Wang, Pavan Kapanipathi, Ryan Musa, Mo Yu, Kartik Talamadupula, Ibrahim Abdelaziz, Maria Chang, Achille Fokoue, Bassem Makni,

Nicholas Mattei, et al. 2019. Improving natural language inference using external knowledge in the science questions domain. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7208–7215.

Yichong Xu, Chenguang Zhu, Ruochen Xu, Yang Liu, Michael Zeng, and Xuedong Huang. 2020. Fusing context into knowledge graph for commonsense reasoning. *arXiv preprint arXiv:2012.04808*.

Vikas Yadav, Steven Bethard, and Mihai Surdeanu. 2019. Quick and (not so) dirty: Unsupervised selection of justification sentences for multi-hop question answering. 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 2578–2589, Hong Kong, China. Association for Computational Linguistics.

Vikas Yadav, Steven Bethard, and Mihai Surdeanu. 2020. Unsupervised alignment-based iterative evidence retrieval for multi-hop question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4514–4525, Online. Association for Computational Linguistics.

Bishan Yang and Tom Mitchell. 2017. Leveraging knowledge bases in LSTMs for improving machine reading. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1436–1446, Vancouver, Canada. Association for Computational Linguistics.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William 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*, pages 2369–2380, Brussels, Belgium. Association for Computational Linguistics.

Wenxuan Zhang, Yang Deng, and Wai Lam. 2020a. Answer ranking for product-related questions via multiple semantic relations modeling. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval, SIGIR 2020*, pages 569–578.

Yao Zhang, Xu Zhang, Jun Wang, Hongru Liang, Adam Jatowt, Wenqiang Lei, and Zhenglu Yang. 2020b. Gmh: A general multi-hop reasoning model for kg completion.

Chen Zhao, Chenyan Xiong, Xin Qian, and Jordan Boyd-Graber. 2020. Complex factoid question answering with a free-text knowledge graph. In *Proceedings of The Web Conference 2020*, pages 1205–1216.

Wanjun Zhong, Jingjing Xu, Duyu Tang, Zenan Xu, Nan Duan, Ming Zhou, Jiahai Wang, and Jian Yin. 2020. Reasoning over semantic-level graph for fact checking. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6170–6180, Online. Association for Computational Linguistics.

Chen Zhu, Yu Cheng, Zhe Gan, Siqi Sun, Tom Goldstein, and Jingjing Liu. 2020. Freelb: Enhanced adversarial training for natural language understanding. In *International Conference on Learning Representations*.

Fengbin Zhu, Wenqiang Lei, Youcheng Huang, Chao Wang, Shuo Zhang, Jiancheng Lv, Fuli Feng, and Tat-Seng Chua. 2021a. Tat-qa: A question answering benchmark on a hybrid of tabular and textual content in finance.

Fengbin Zhu, Wenqiang Lei, Chao Wang, Jianming Zheng, Soujanya Poria, and Tat-Seng Chua. 2021b. Retrieving and reading: A comprehensive survey on open-domain question answering. *arXiv preprint arXiv:2101.00774*.

A Appendix

A.1 Case Study

More case studies can be found in Table 7.

(a) Case Study 1

Question:	*Algae can be found in* **(A) reservoir** (B) meat (C) street (D) tree
Useful facts retrieved by IR:	Algae is found in bodies of water.
Additional facts from path-based analytics:	Water reservoir: a large quantity of water is stored in a reservoir (or dam).
Relevant path in AMR-SG:	Algae → find-01 → body → water → store-01 → reservoir

(b) Case Study 2

Question:	*Photosynthesis can be performed by* **(A) a cabbage cell** (B) a bee cell (C) a bear cell (D) a cat cell
Useful facts retrieved by IR:	N.A.
Additional facts from path-based analytics:	Plant cells can perform photosynthesis. Description: skunk cabbage is a flowering perennial plant that is one of the first plants to emerge in the spring.
Relevant path in AMR-SG:	Photosynthesis → plant → cabbage

(c) Case Study 3

Question:	*Which is recyclable?* (A) An Elephant **(B) A school notebook** (C) A boat (D) A lake
Useful facts retrieved by IR:	Paper is recyclable.
Additional facts from path-based analytics:	Take notes on notebook paper.
Relevant path in AMR-SG:	recycle-01 → paper → notebook

(d) Case Study 4

Question:	*Which requires energy to move?* **(A) weasel** (B) willow (C) mango (D) poison ivy
Useful facts retrieved by IR:	An animal requires energy to move.
Additional facts from path-based analytics:	The long and slender body of the weasel allows it to move, almost flow, over terrain.
Relevant path in AMR-SG:	energy → move-01 → weasel

(e) Case Study 5

Question:	*A person wants to be able to have more natural power in their home. They choose to cease using a traditional electric company to source this electricity, and so decide to install* (A) sun grafts (B) sunlight shields **(C) panels collecting sunlight** (D) solar bees
Useful facts retrieved by IR:	A home with solar electric panels on the roof might be able to make most of its own electricity, for example.
Additional facts from path-based analytics:	Solar thermal panels generate hot water from the natural energy in sunlight.
Relevant path in AMR-SG:	natural-03 → energy → generate-01 → sunlight → panel

Table 7: More case studies in addition to Table 6