2022/04/22 ゼミ

朱羿華

論文紹介

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

Xu W, Zhang H, Cai D, et al. Dynamic semantic graph construction and reasoning for explainable multi-hop science question answering[J]. arXiv preprint arXiv:2105.11776, 2021.

ACL 2021

朱羿華(シュ　ゲイカ)

チエック担当者: 内藤 雅博

下平研究室ゼミ

2022.04.22

Introduction

Requirements for Multi-hop QA:

1, collection of information from large external knowledge resources
2, aggregation of retrieved facts to answer complex natural language questions

External knowledge for QA

	Textual corpora	graph structure
Advantages	1, contain rich and diverse **evidence facts** 2, the success of pre-trained models (LMs)	provide structural clues about relevant entities for explainable predictions
Disadvantages	Cannot retrieve relevant and useful facts to fill the **knowledge gap** for inferring the answer.	suffer from sparsity, where complex question clues are unlikely to be covered by the closed-form relations in KG

Introduction

Question: Predators eat __.
Answer Choice: bunnies
Hypothesis: Predators eat bunnies.

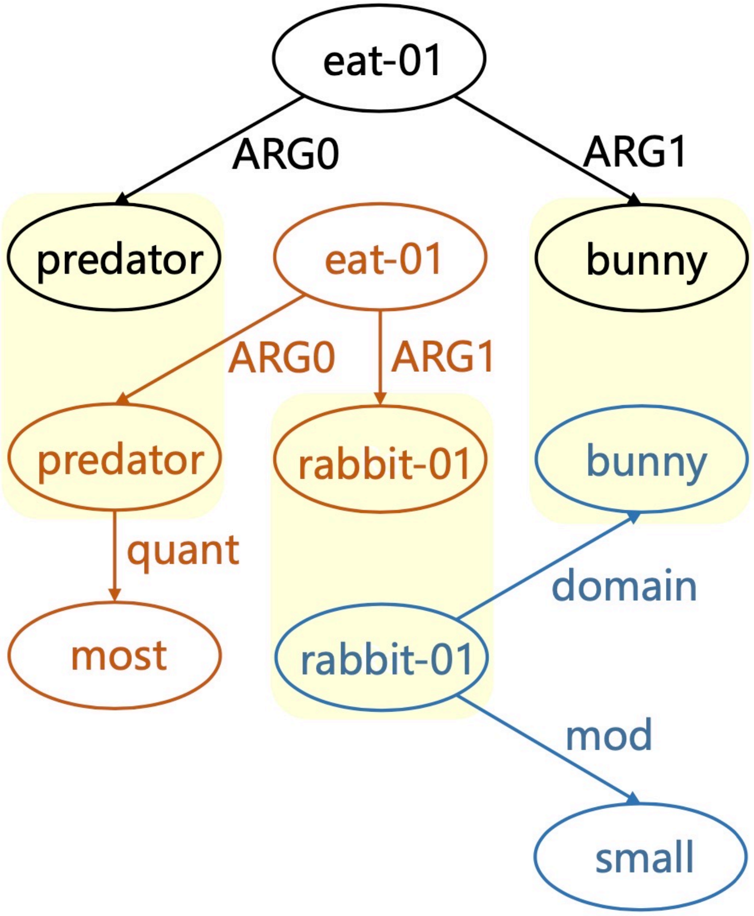
Fact 1: A bunny is a small rabbit.
Fact 2: Most predators eat rabbits.

How to utilize all External Knowledge?

Use **Abstract Meaning Representation (AMR)** as a graph annotation to a textual fact

Definition:
AMR is a semantic formalism that represents the meaning of a sentence into a rooted, directed graph.

Target:
The aim of AMR is to capture every meaningful content in high-level abstraction while removing away inflections and function words in a sentence.



AMR example

Introduction

In this paper

Novel methods applied in this paper: (Will be introduced in the approach part)

1, AMR-SG (AMR-based Semantic Graph)
2, A novel path-based fact analytics approach exploiting AMR-SG
3, A fact-level relation modeling leveraging GCN

Experimental results: (outperform previous approaches that use additional KGs)

1, OpenBookQA: 81.6
2, ARC-Challenge: 68.94

Approach (Overview)

1, Fact Retrieval

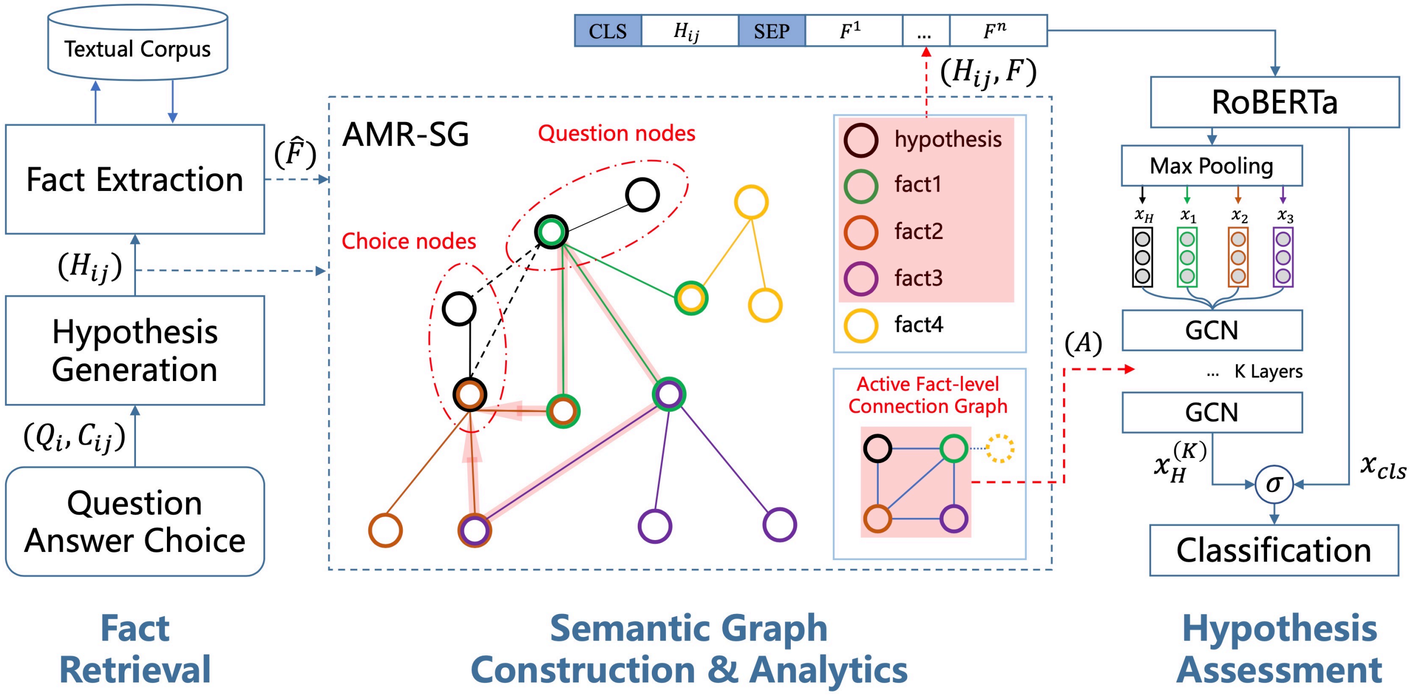
retrieve evidence facts $\hat{F} = \{\hat{F}^1, \dots, \hat{F}^m\}$ for each question-choice pair from a large textual corpus.

2, Semantic Graph Construction & Analytics

dynamically constructs a semantic graph, named AMR-SG, to select active facts $\hat{F} = \{\hat{F}^1, \dots, \hat{F}^m\}$ from $\hat{F}$ and capture their relations A.

3, Hypothesis Assessment

classifies whether the question-choice is correct, given the active facts and their relations in (2).



Q_i : i^{th} Question.
C_{ij}, j $\in \{1,2,\dots,J\}$: Answer choices j in the Question i.
H_{ij} : Hypothesis for the i^{th} Question , j^{th} Answer.
x_i : vector representation for i^{th} fact.

Approach (1, Fact Retrieval)

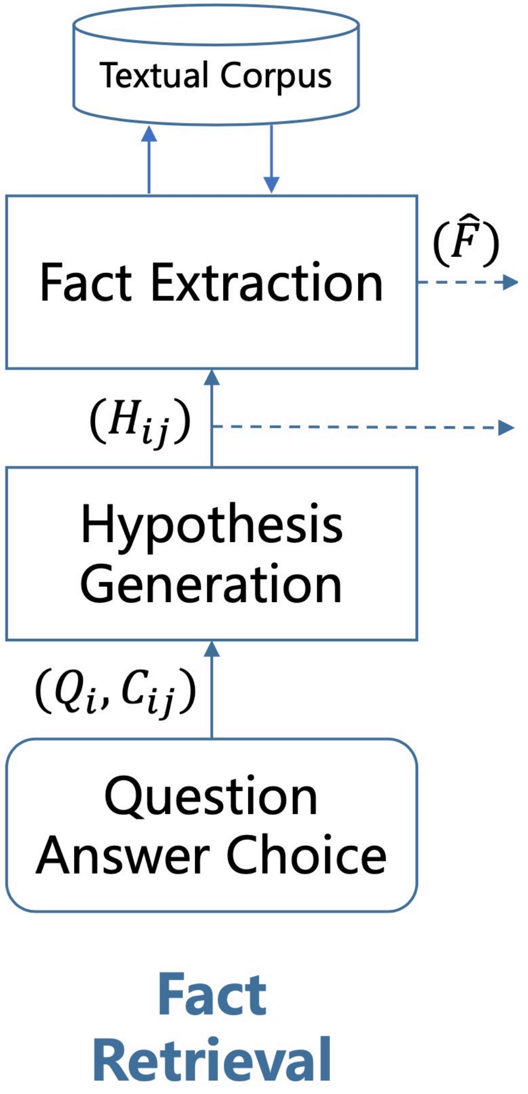
Hypothesis Generation:

A hypothesis is a completed statement derived from each question-choice pair

Fact Extraction:

retrieve a pool of evidence facts $\hat{F}$ for each hypothesis

Algorithm: ***Elasticsearch*** (Gormley and Tong, 2015)



Approach

(2, Semantic Graph Construction & Analytics)

2.1, AMR-SG Construction

Basic information:

AMR structure algorithm: AMR parser (Cai and Lam, 2020)

AMR $G = \{G^H, G^1, \ldots, G^{\mathrm{m}}\}$

G^H : hypothesis G^i : i^{th} fact

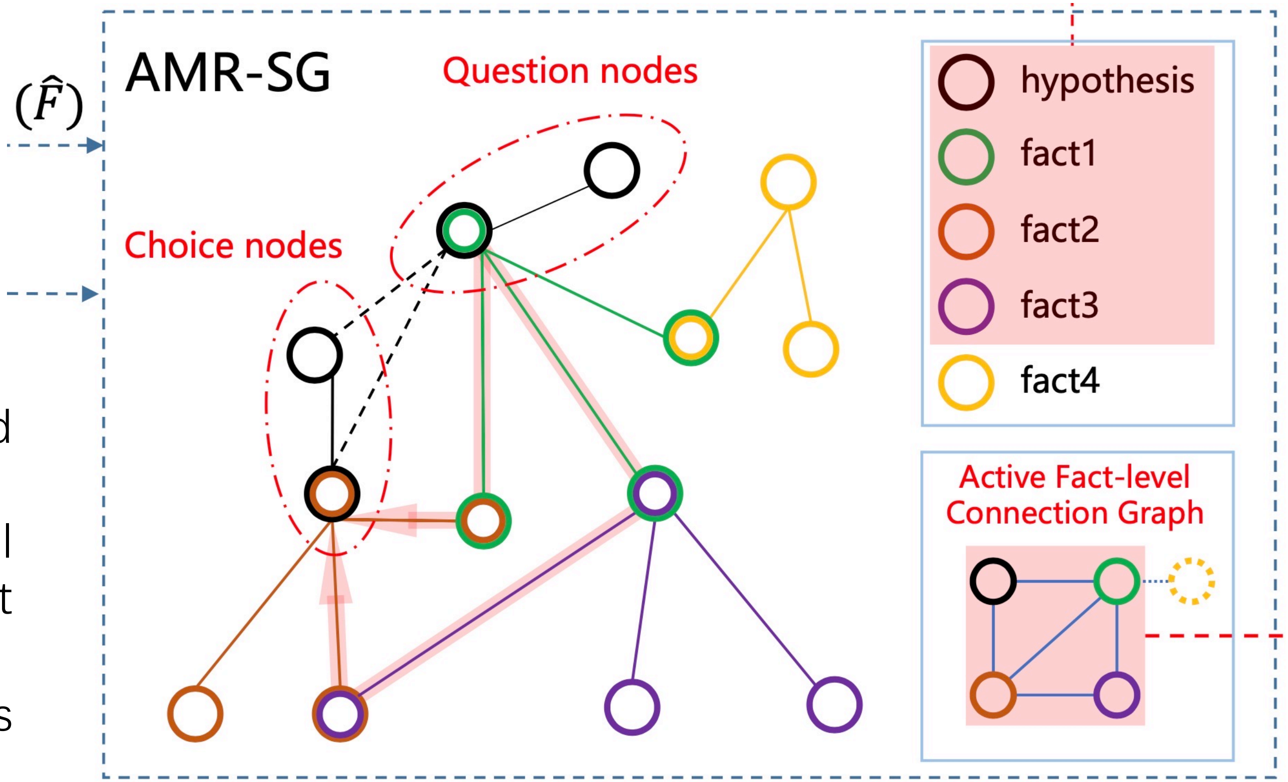
Construction:

1, start from G^H. (G^H contains question nodes Q^H and choice nodes C^H)
2, 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.
3, stop when no AMR can be added or the fact pool is empty.

Definition:

$$Q_{ij}^H = \cap_{j=1}^{J} \{v \mid v \in G_{ij}^H\}$$

$$C_{ij}^H = \{v \mid v \in G_{ij}^H, v \notin Q_{ij}^H\}, j = 1, \ldots, J$$



Approach

(2, Semantic Graph Construction & Analytics)

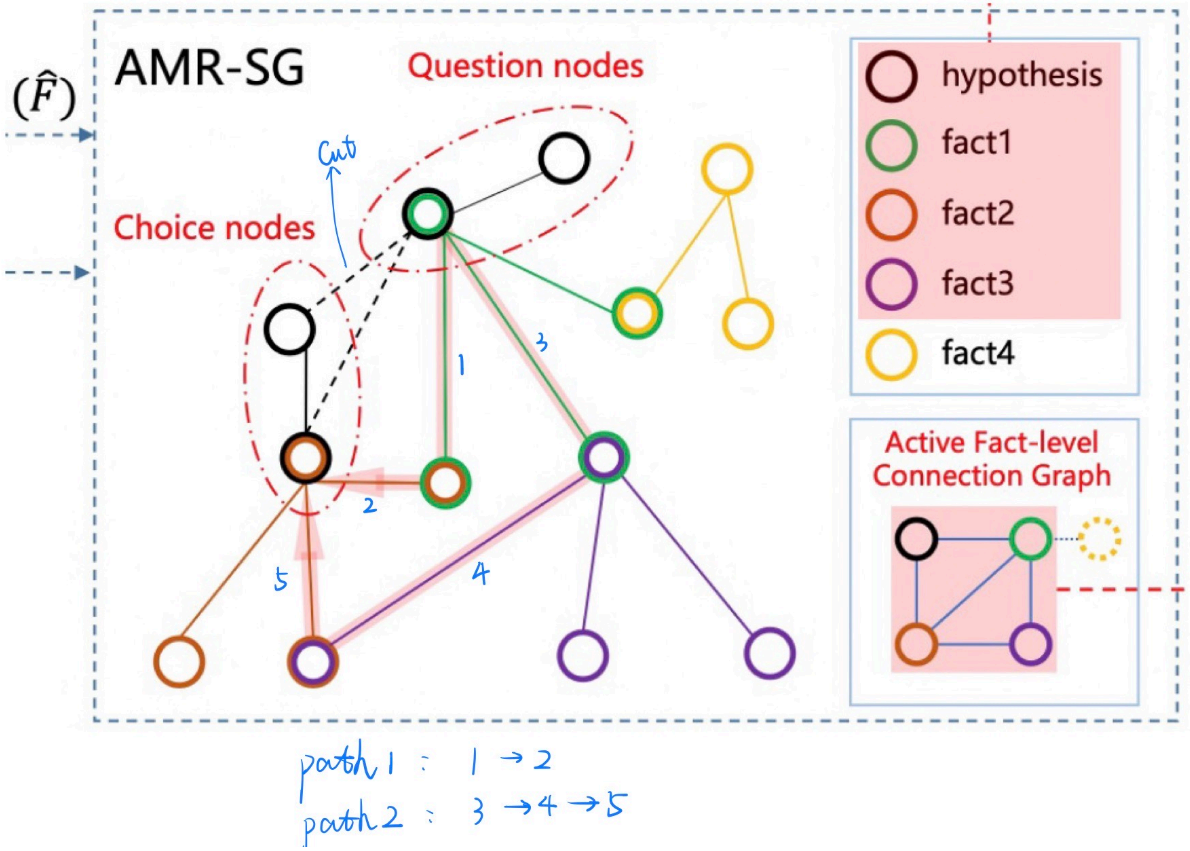
2.2, Path-based Analytics

1, **cut** the edges between G^H and C^H to guarantee the paths are spotted outside G^H.

2, apply **depth-first search** on AMR-SG to find all paths that connect at least one question node and one choice node . (e.g. path1 and path2)

3, abandon excess facts (e.g.Fact 4).

4, construct an **Active Fact-level Connection Graph** from AMR-SG to capture such relations among the hypothesis and all active facts



Approach (3, Hypothesis Assessment with Fact-level Reasoning)

1, concatenate the hypothesis and all active facts to RoBERTa to get the hidden representations of the hypothesis($s^H_{1:l_H} \in \mathbb{R}^{L_H \times d}$) and the i^{th} active facts($s^i_{1:l_i} \in \mathbb{R}^{L_i \times d}$).

2, A max pooling layer is applied over these hidden representations to get the node representations respectively.

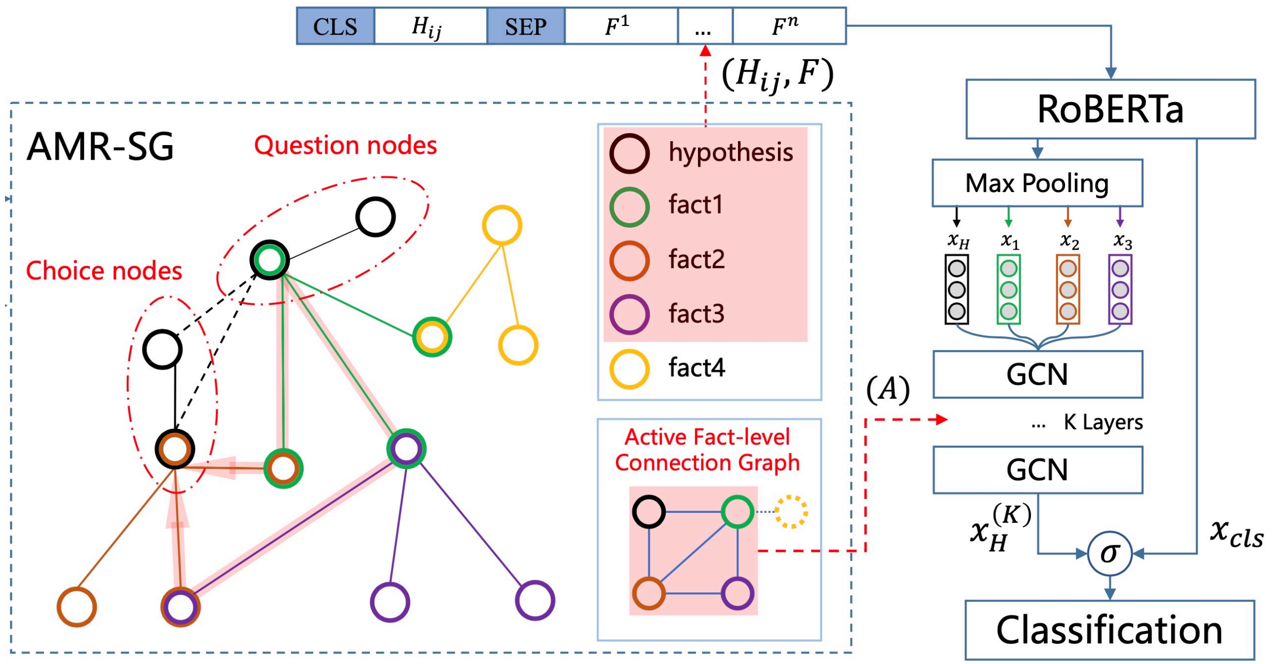
$$x_H = MaxPool\left(s^H_{1:l_H}\right) \in \mathbb{R}^{1\times d}$$
$$x_i = MaxPool\left(s^i_{1:l_i}\right) \in \mathbb{R}^{1\times d}, i = 1, \ldots, n$$

3, construct the AFCG:

node: x_H, x_i

Edge: simple adjacency matrix $A \in \mathbb{R}^{(n+1)\times(n+1)}$.

$$A_{ij} = \begin{cases} 1, if\ F^i\ is\ connected\ with\ F^j \\ 0, \quad otherwise \end{cases}$$



Approach

(3, Hypothesis Assessment with Fact-level Reasoning)

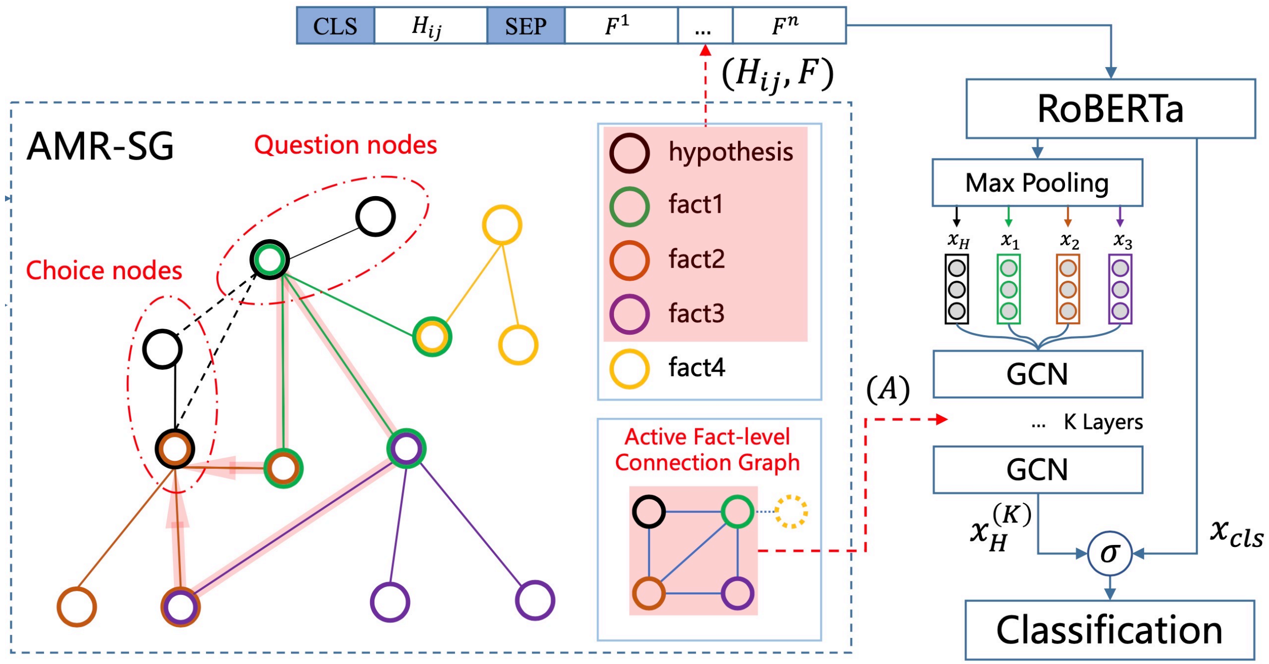
4, use GCN(graph convolutional network) to do the representation learning. (K layers)

5, use final layer hypothesis representation $x_H^{(K)}$ and x_{cls} to get the final probability.

$$\lambda = \sigma\left(W^{\lambda}\left[x_{cls}: x_H^{(K)}\right] + b^{\lambda}\right)$$

$$s(q,a) = W^{0}\left(\lambda x_H^{(K)} + (1-\lambda)x_{cls}\right) + b^{0}$$

W^{λ}, b^{λ}, W^{0}, b^{0} are learnable parameters.



Experiment

Dataset:

1, multi-hop QA datasets:

ARC-Challenge (Clark et al., 2018)
OpenBookQA (Mihaylov et al., 2018)

2, textual corpus:

ARC Corpus (Clark et al., 2018)

Implementation:

1, OpenBookQA:

learning rate: $2\times\mathbf{10^{-5}}$
batch size: 12

2, ARC-Challenge:

learning rate: $1\times\mathbf{10^{-5}}$
batch size: 6

3, GCN layers: 2

Results

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.

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.

OpenBookQA

ARC-Challenge

Explainability Analysis (1, Impact of Evidence Facts)

Facts Composition (total 15 facts)	Core Fact Retrieval Accuracy	Human Evaluation Rel.	Info.	Comp.	Test set Accuracy 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.

Score for Relatedness and Informativeness:

One fact contributes 1 score if it meets the requirement of Relatedness and Informativeness.

Score for Completeness:

all 15 facts contribute 1 score if they together meet the requirement of Completeness

Core fact: the facts from open-book.
Common fact: facts from ARC Corpus.
Simple IR system: simple information retrieval (IR) system (*Elasticsearch*)

For core facts, open-book annotates **one gold fact** for each question, therefore, we evaluate the quality by using the **retrieval accuracy** of the gold fact.

For common facts, only can use human analysis:

1. Relatedness: Does the retrieved fact related to the question or the answer?
2, Informativeness: Does the retrieved fact provide useful information to answer the question?
3, Completeness: Do all retrieved facts together fill the knowledge gap to completely answer the question?

Explainability Analysis (1, Impact of Evidence Facts)

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.

Analysis 1: our approach makes an overall improvement with regard to Relatedness, Informativeness and Completeness. (**Use "fact" more effectively**)

Analysis 2: 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.**

Explainability Analysis (2, case study)

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.

Cannot retrieve useful facts by IR system.

Can form a complete reasoning chain by AMR-SG

References

Xu W, Zhang H, Cai D, et al. Dynamic semantic graph construction and reasoning for explainable multi-hop science question answering[J]. arXiv preprint arXiv:2105.11776, 2021.

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

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

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

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