Connecting the Dots:
A Knowledgeable Path Generator for Commonsense Question Answering

Peifeng Wang Nanyun Peng Filip Illievski Pedro Szekely Xiang Ren

Department of Computer Science, University of Southern California
Department of Computer Science, University of California, Los Angeles
Information Sciences Institute, University of Southern California

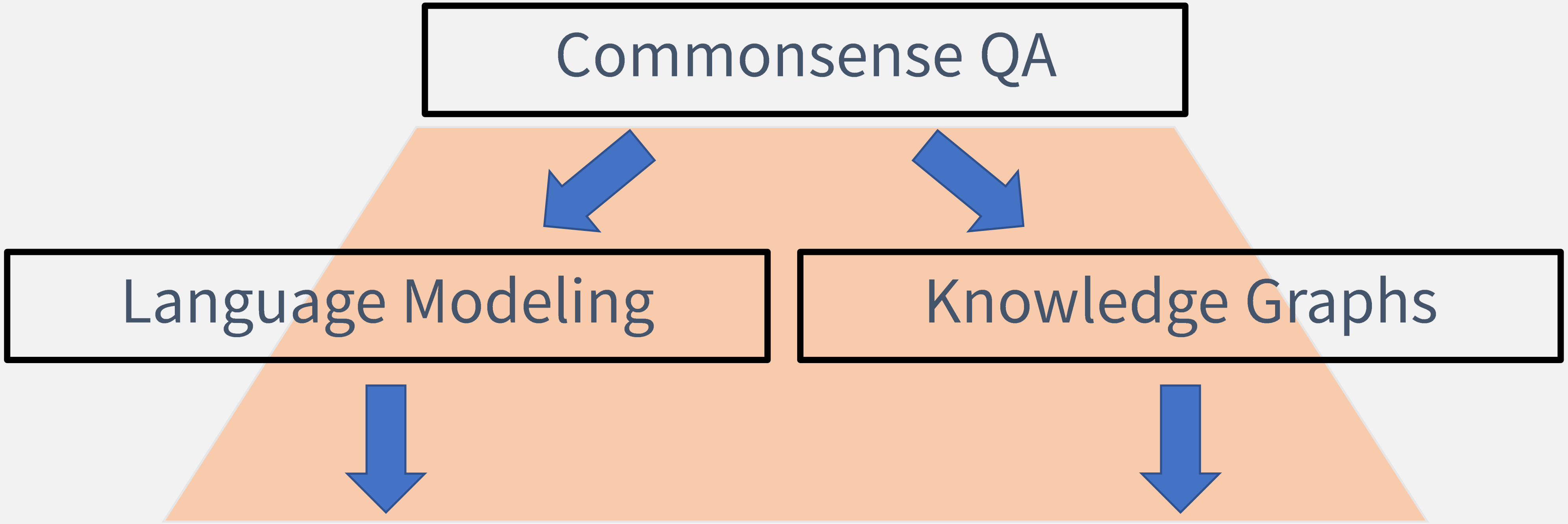
Noah Lee
noahlee357@korea.ac.kr
Data Intelligence Lab, Korea University
2020 01. 15

EMNLP ’20

01 Problem Statement

How can we empower the Knowledge Graphs(KG) for Commonsense QA?

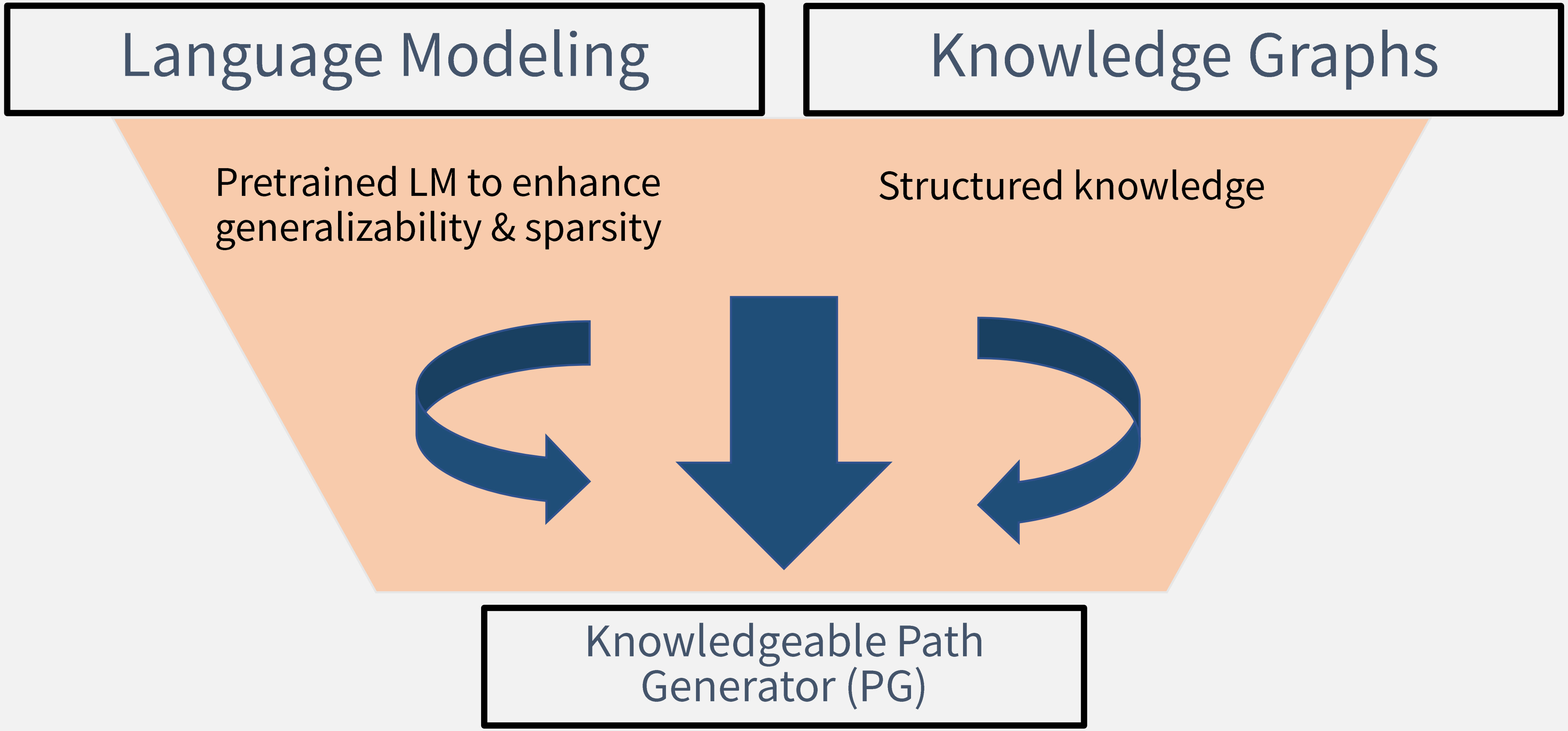
Motivation



- Commonsense reasoning vs Spurious correlation
- Uninterpretable

- Sparsity Problem
- Contextualization Problem

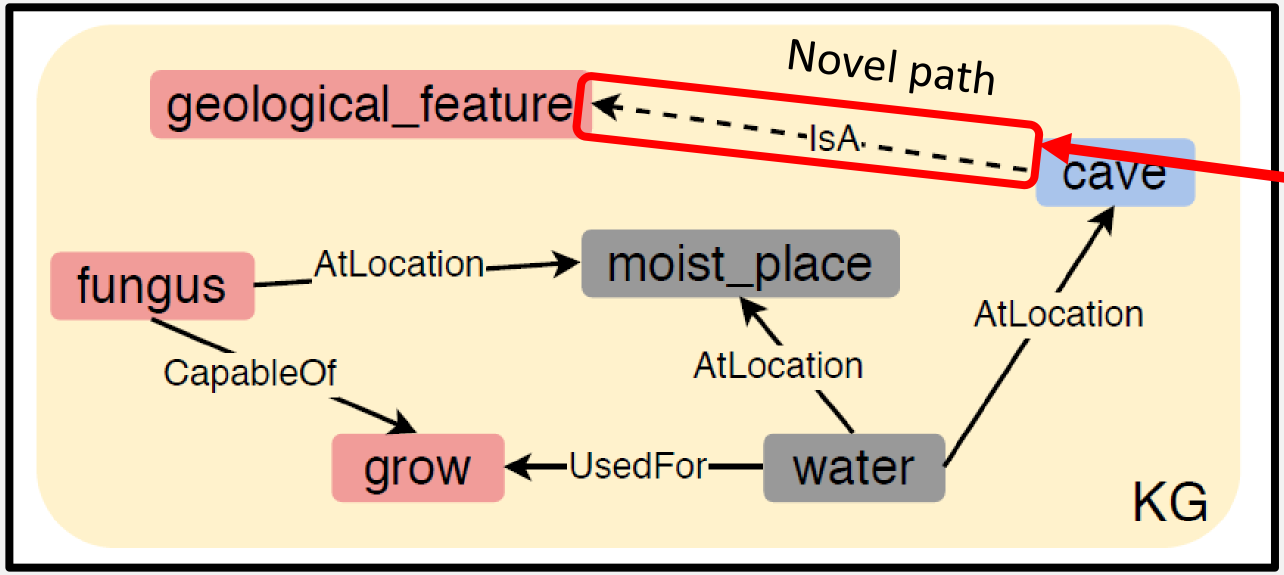
02 Motivation



02 Motivation

Q: In what geological feature will you find fungus growing?

A: shower stall B: toenails C: basement D: forest E: cave



Knowledgeable Path Generator (PG)

03 Contribution

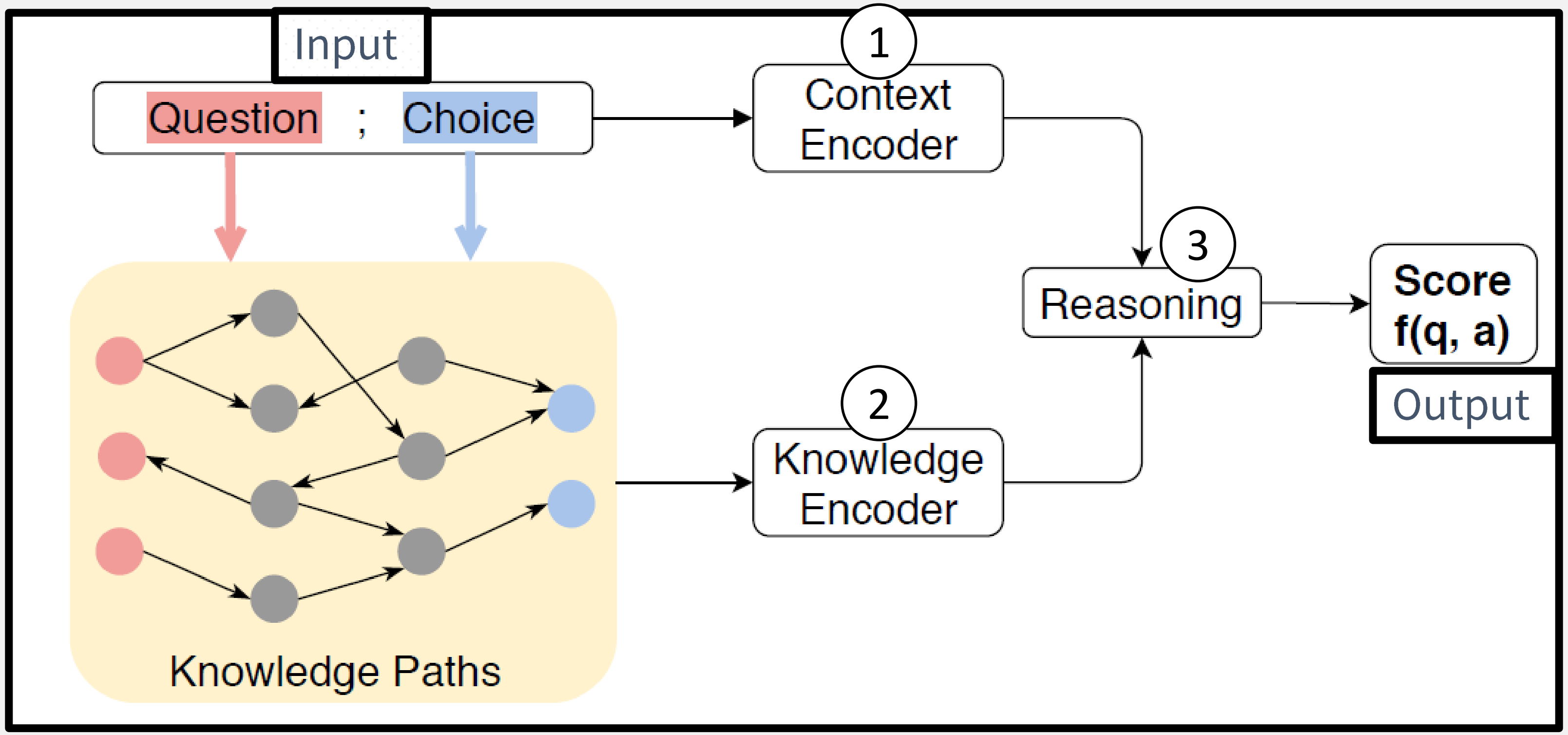
1) Proposes a method to generate task-relevant knowledge paths that may not **exist** in the original KG

2) Design and implement a QA framework with the **Knowledgeable Path Generator(PG)**

3) Demonstrates **effectiveness** of the method by extensive experiments on **two benchmark datasets** (CQA & OBQA)

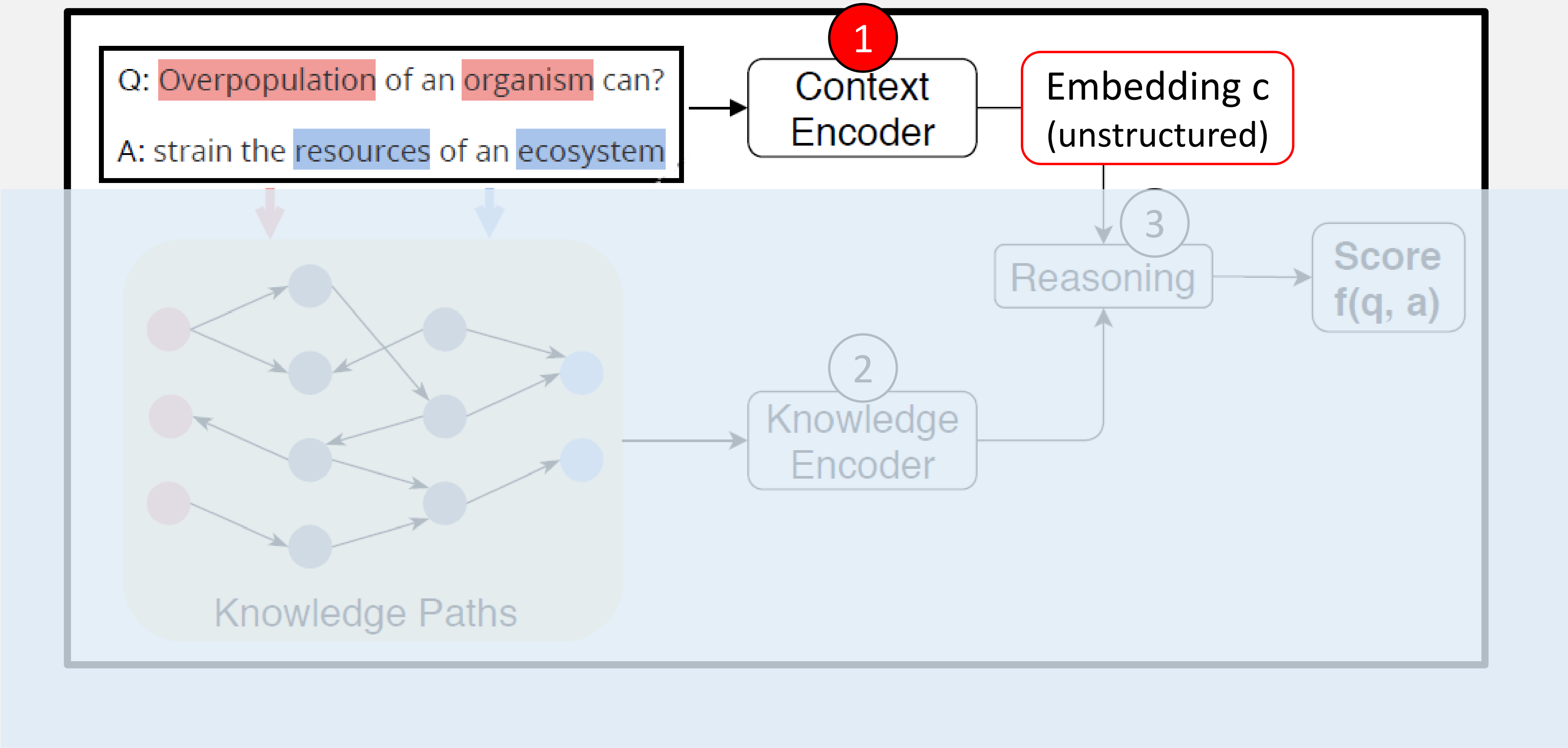
04 Model

KG-augmented QA Framework



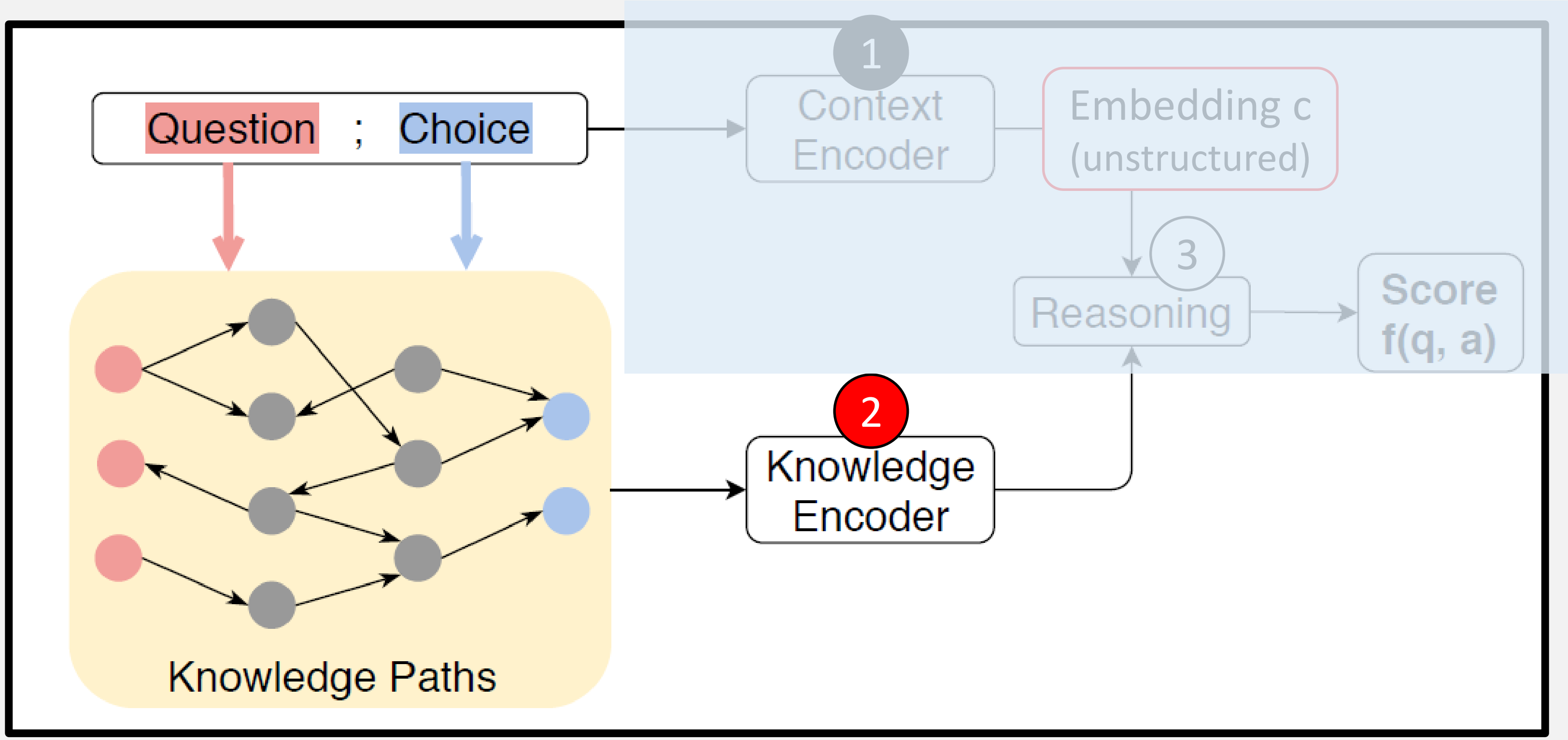
04 Model

KG-augmented QA Framework: Context Module

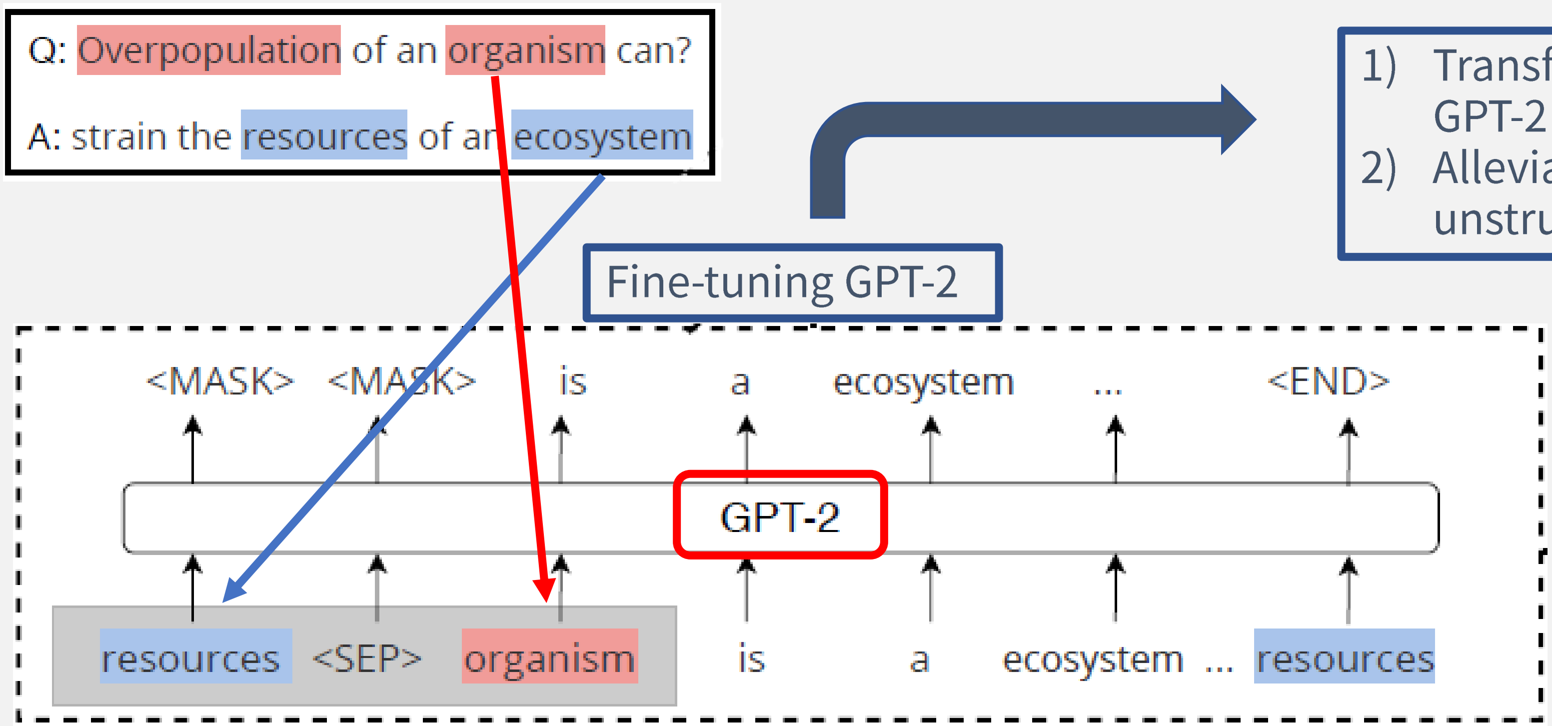


04 Model

KG-augmented QA Framework: Knowledge Module

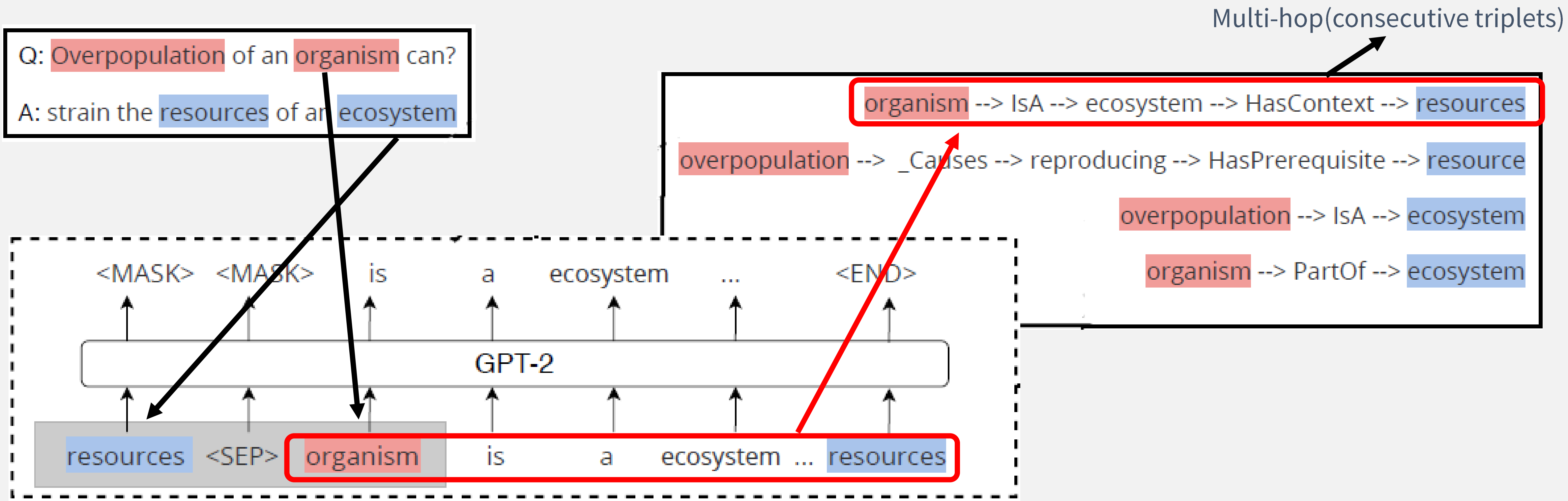


KG-augmented QA Framework: Knowledge Module – Path Generator



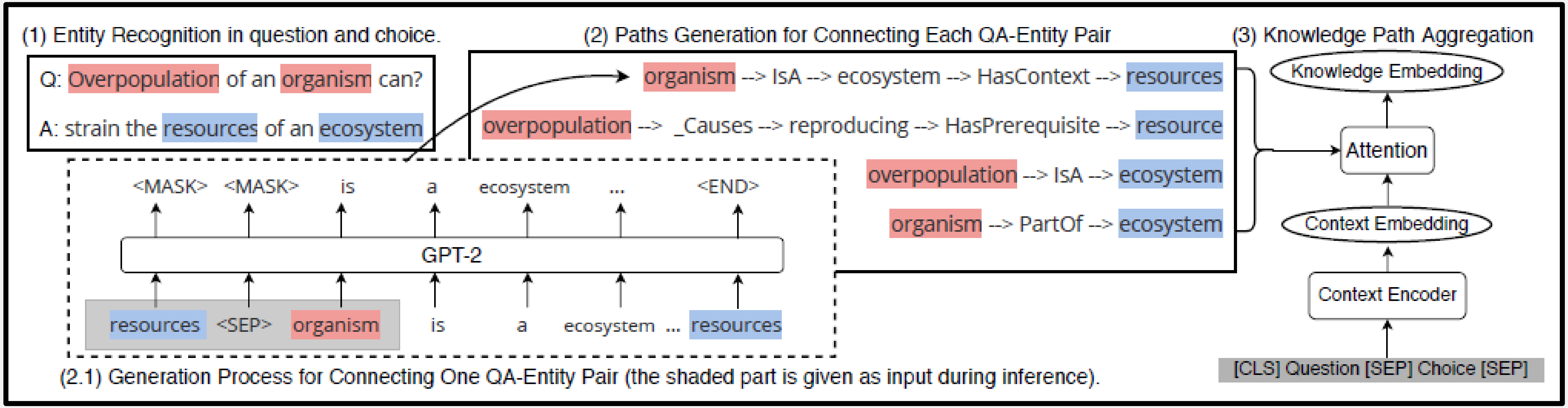
1) Transfer knowledge of pre-trained GPT-2 to PG
2) Alleviate sparsity problem with unstructured knowledge in GPT-2

KG-augmented QA Framework: Knowledge Module – Path Generator

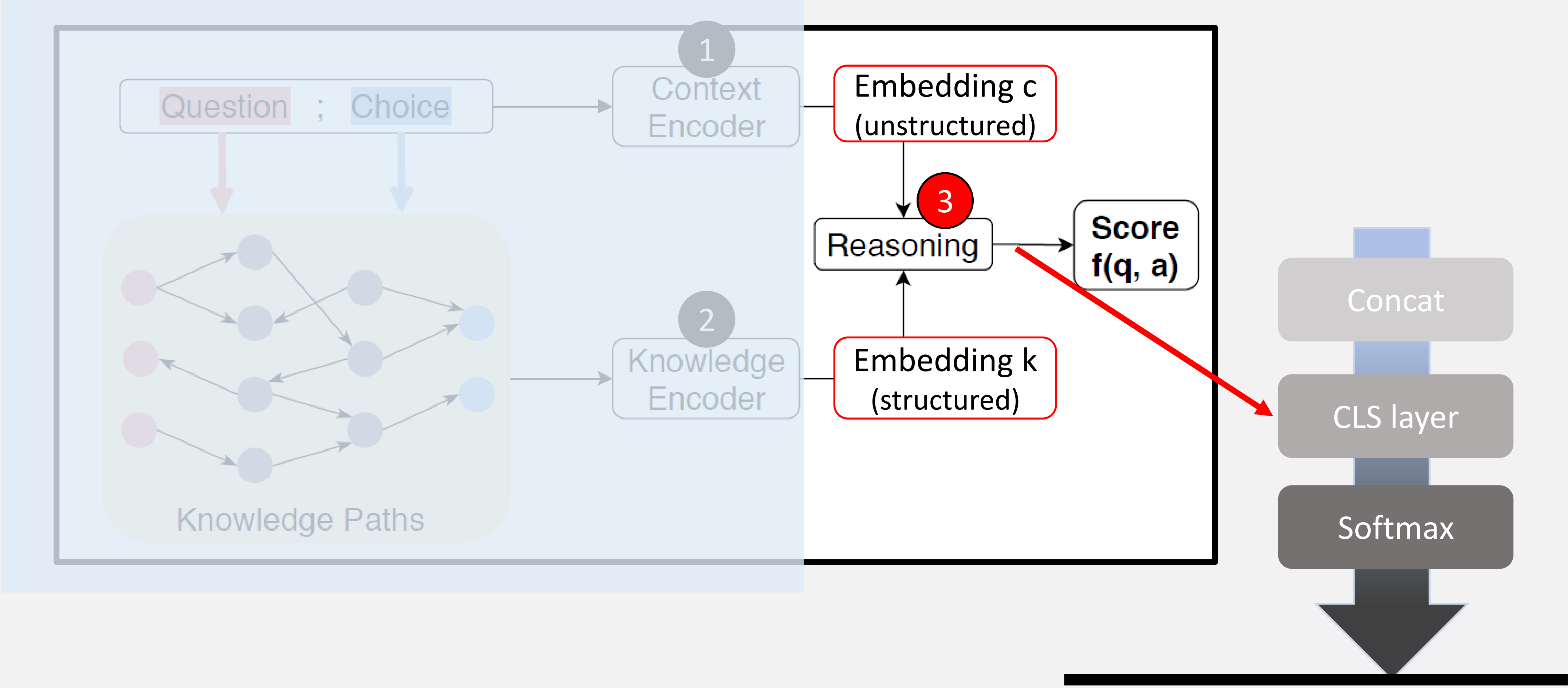


04 Model

KG-augmented QA Framework: Knowledge Module



KG-augmented QA Framework: Reasoning Module



05 Datasets / Code / Settings

[Dataset]

Commonsense QA [Test set unavailable]

https://www.tau-nlp.org/commonsenseqa

Openbook QA

https://leaderboard.allenai.org/open_book_qa/submissions/public

[Code]

https://github.com/wangpf3/Commonsense-Path-Generator

05 Datasets / Code / Settings

[Entity Recognition]

– Conceptnet(Speer et al., 2017)

[Path Sampling]

- Hops: 1~3
- Global Sampling: 2.8 mil paths
- Local Sampling: 133k paths(CQA)
  105k paths(OBQA)
- Split: 90:5:5 (train/dev/test)

[Path Generator Training]

– Fine-tune GPT-2
- LR: 1e-5
- Batch size: 64
- Early Stopping : 2 epochs of consecutive loss

(Context Module Settings in Appendix)

06

Baselines

Fine-tuned LM (w/o KG)

Static KG Models

1) Relational Network (RN)
2) Relational Graph Convolutional Network (RGCN)
3) GconAttn

Link Prediction Model – 1 hop path embedding

07 Experiments

OBQA Test Accuracy

Methods	RoBERTa-large	AristoRoBERTa
Fine-tuned LMs (w/o KG)	64.80 (±2.37)	78.40 (±1.64)
+ RN	65.20 (±1.18)	75.35 (±1.39)
+ RGCN	62.45 (±1.57)	74.60 (±2.53)
+ GconAtten	64.75 (±1.48)	71.80 (±1.21)
+ Link Prediction	66.30 (±0.48)	77.25 (±1.11)
+ PG-Local	70.05 (±1.33)	79.80 (±1.45)
+ PG-Global	68.40 (±0.31)	**80.05** (±0.68)
+ PG-Full	**71.20** (±0.96)	79.15 (±0.78)

Commonsense QA Test Accuracy

Methods	Single	Ensemble
RoBERTa (Liu et al., 2019)	72.1	72.5
RoBERTa+FreeLB (Zhu et al., 2019)	-	73.1
RoBERTa+HyKAS (Ma et al., 2019)	73.2	-
XLNet+DREAM	73.3	-
RoBERTa+KE	-	73.3
RoBERTa+KEDGN	-	74.4
XLNet+GraphReason (Lv et al., 2019)	75.3	-
Albert (Lan et al., 2019)	-	76.5
UnifiedQA* (Khashabi et al., 2020)	79.1	-
Albert+PG-Full	75.6	78.2

07 Experiments

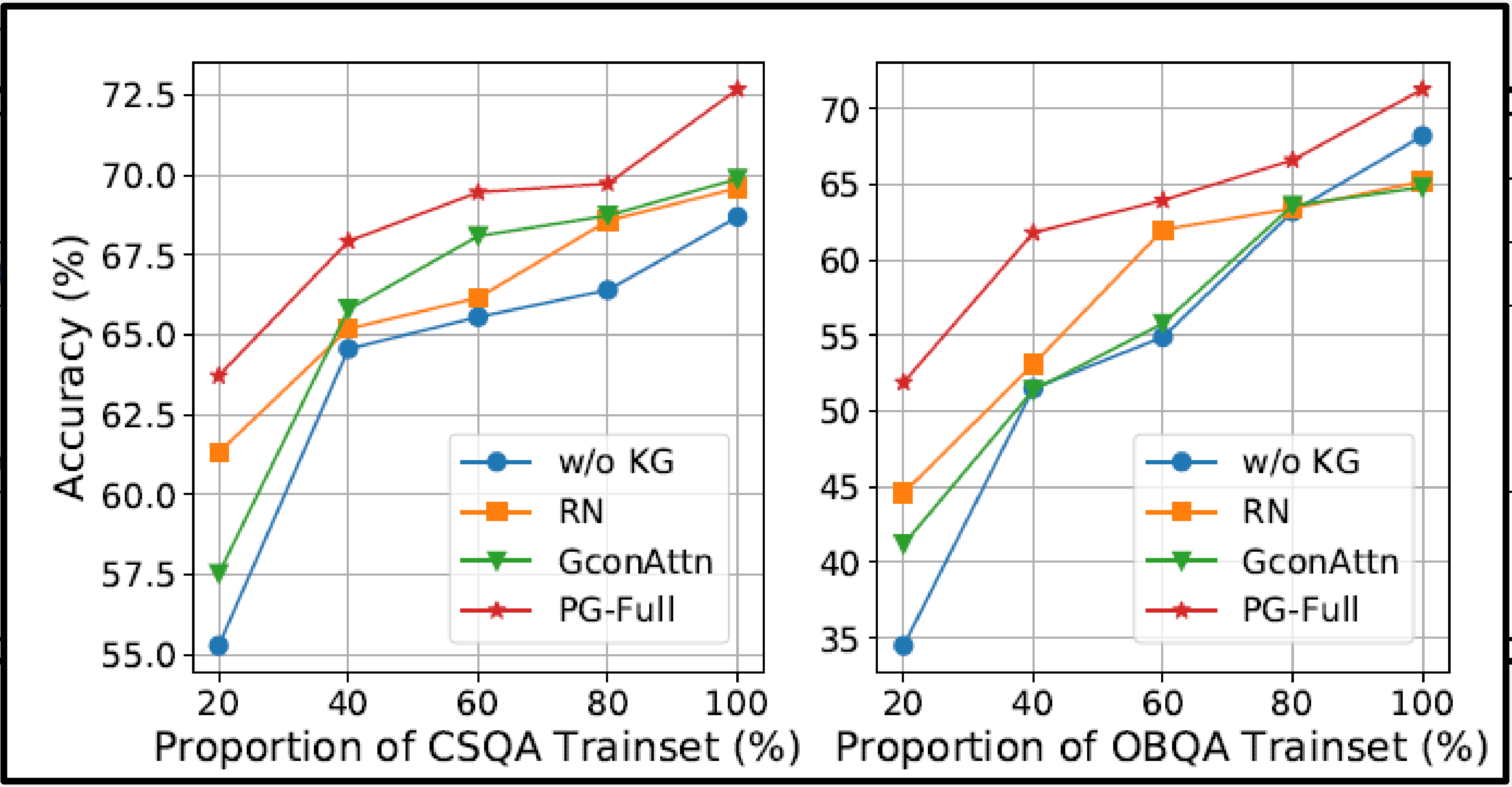
Commonsense QA Test Accuracy with varying proportion of train data

Methods	BERT-large			RoBERTa-large		
	20% Train	60% Train	100% Train	20% Train	60% Train	100% Train
Fine-tuned LM (w/o KG)	46.25 (±0.63)	52.30 (±0.16)	55.39 (±0.40)	55.28 (±0.35)	65.56 (±0.76)	68.69 (±0.56)
+ RN	45.12 (±0.69)	54.23 (±0.28)	58.92 (±0.14)	61.32 (±0.68)	66.16 (±0.28)	69.59 (±3.80)
+ RGCN	48.67 (±0.28)	54.71 (±0.37)	57.13 (±0.36)	58.58 (±0.17)	68.33 (±0.85)	68.41 (±0.66)
+ GconAttn	47.95 (±0.11)	54.96 (±0.69)	56.94 (±0.77)	57.53 (±0.31)	68.09 (±0.63)	69.88 (±0.47)
+ Link Prediction	47.10 (±0.79)	53.96 (±0.56)	56.02 (±0.55)	60.84 (±1.36)	66.29 (±0.29)	69.33 (±0.98)
+ PG-Local	50.20 (±0.31)	55.68 (±0.07)	56.81 (±0.73)	61.56 (±0.72)	67.77 (±0.83)	70.43 (±0.65)
+ PG-Global	49.89 (±1.03)	55.47 (±0.92)	57.21 (±0.45)	62.93 (±0.82)	68.65 (±0.02)	71.55 (±0.99)
+ PG-Full	**51.97** (±0.26)	**57.53** (±0.19)	**59.07** (±0.30)	**63.72** (±0.77)	**69.46** (±0.23)	**72.68** (±0.42)

07 Experiments

Common

Methods	100% Train
Fine-tuned LM (	68.69 (±0.56)
+ RN	69.59 (±3.80)
+ RGCN	68.41 (±0.66)
+ GconAttn	69.88 (±0.47)
+ Link Predictio	69.33 (±0.98)
+ PG-Local	70.43 (±0.65)
+ PG-Global	71.55 (±0.99)
+ PG-Full	**72.68** (±0.42)



Ablation Study

Methods	CQA	OBQA
Scratch (fine-tune X)	68.75	65.50
Full (fine-tune)	72.68	71.20

1) Scratch approximates what static KG already has

2) Pre-trained GPT-2 complements missing knowledge of a static KG

08 Limitation

Model	Affiliation	Date	Accuracy	Accuracy (*Uses ConceptNet)
Human		03/10/2019	88.9	
ALBERT+DESC-KCR (ensemble model)	Microsoft Cognitive Services Research	12/02/2020		83.3
Albert+KD (ensemble model)	HIT-SCIR-QA	12/30/2020		80.9
ALBERT+DESC-KCR (single model)	Microsoft Cognitive Services Research	12/02/2020		80.7
ALBERT+KD (single model)	HIT-SCIR-QA	12/10/2020		80.3
Albert + KCR(knowledge chosen by relations, single model)	ITNLP (Harbin Institute of Technology)	07/12/2020		79.5
UnifiedQA (single model)	Allen Institute for AI	04/23/2020	79.1	
Albert + KRD(single model)	SudaNLP	12/04/2020	78.4	
Albert + PathGenerator (ensemble model)	USC MOWGLI / INK Lab	05/14/2020		78.2
T5 (single model)	Allen Institute for AI	04/23/2020	78.1	
TeGBERT (single model)	anonymous	07/22/2020		76.8
ALBERT (ensemble model)	Zhiyan Technology	12/18/2019	76.5	

08 Conclusion

1) Addresses the **contextualization** and **sparsity** challenges of original KGs

2) Design and implement a QA framework with the **Knowledgeable Path Generator(PG)**

3) Demonstrates effectiveness of the method in its **robustness** to limited training data.

Q&A