Knowledge-aware Commonsense Question Answering

知识注意的常识问答

——刘平生

Scalable Multi-Hop Relational Reasoning for Knowledge-Aware Question Answering

Yanlin Feng[♣*] **Xinyue Chen**[♠*] **Bill Yuchen Lin**[♥] **Peifeng Wang**[♥] **Jun Yan**[♥] **Xiang Ren**[♥]
fengyanlin@pku.edu.cn, xinyuech@andrew.cmu.edu,
{yuchen.lin, peifengw, yanjun, xiangren}@usc.edu
[♥]University of Southern California
[♣]Peking University [♠]Carnegie Mellon University

EMNLP 2020
https://www.aclweb.org/anthology/2020.emnlp-main.99.pdf

问题描述

Where does a child likely sit at a desk?

A. Schoolroom * B. Furniture store C. Patio
D. Office building E. Library

常识： “a child is likely to appear in a schoolroom”

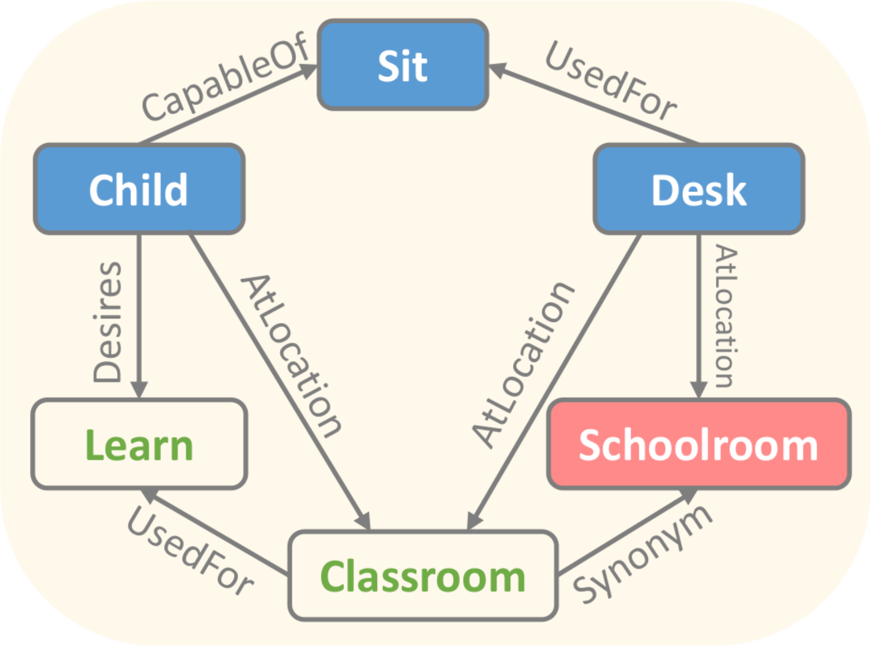
Q: In what geological feature will you find fungus growing?
A: shower stall B: toenails C: basement D: forest E: cave

常识： “cave has the geological feature”

“cave is usually moist”

“fungus grows in moist place”

问题描述



Where does a child likely sit at a desk?

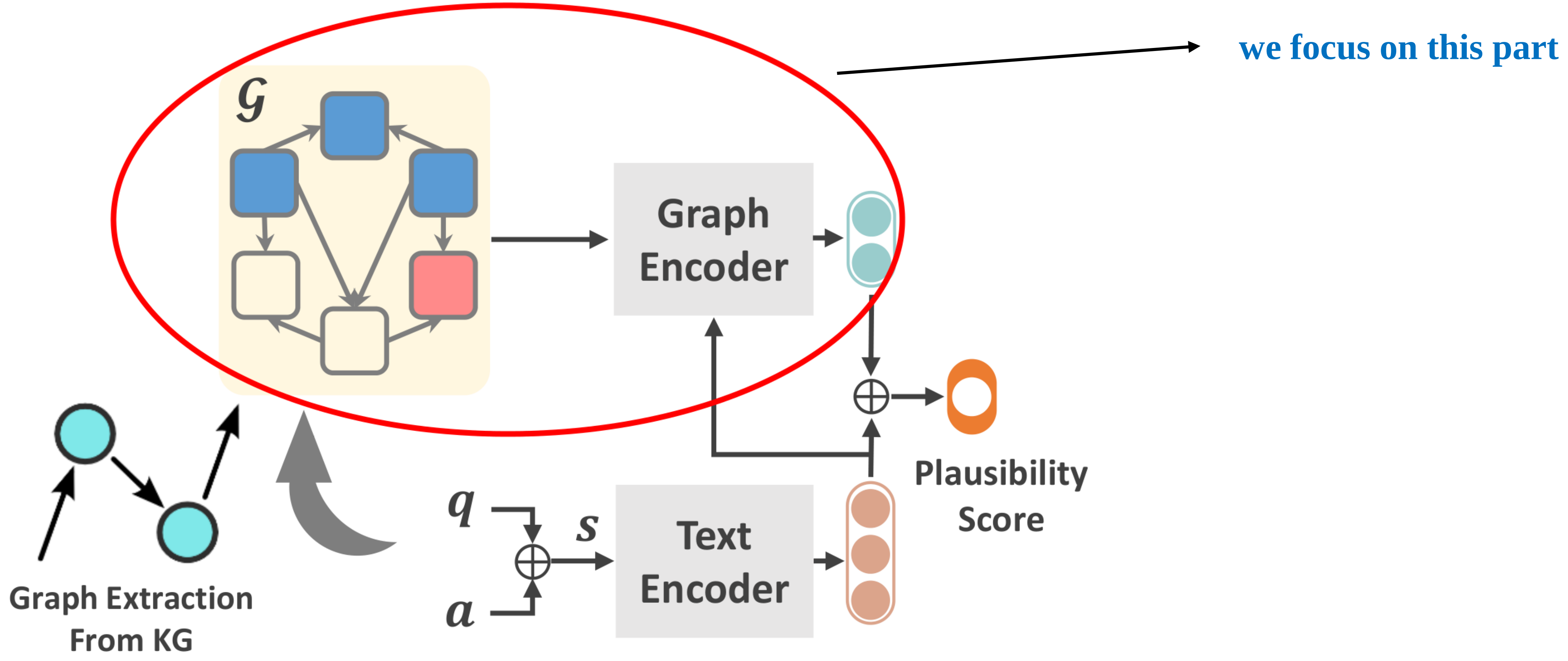
A. Schoolroom * B. Furniture store C. Patio
D. Office building E. Library

Q和A中的实体：{Child, Sit, Desk, Schoolroom }

关系路径：(Child → AtLocation → Classroom → Synonym → Schoolroom)

如何处理这个关系图?

Overview of the knowledge-aware QA framework



Graph Encoding with Path-Based Models

- Relation Network (RN)
- KagNet

直接从图中抽取关系路径，并用序列模型编码的方法

RN：

$$\mathrm{RN}(\mathcal{G}) = \mathrm{Pool}\Big(\{\mathrm{MLP}(\boldsymbol{h}_j \oplus \boldsymbol{e}_r \oplus \boldsymbol{h}_i) \mid j \in \mathcal{Q}, i \in \mathcal{A}, (j, r, i) \in \mathcal{E}\}\Big).$$

KagNet：

$$\textsc{KagNet}(\mathcal{G}) = \mathrm{Pool}\Big(\{\mathrm{LSTM}(j, r_1, j_1, \ldots, r_k, i) \mid (j, r_1, j_1), \cdots, (j_{k-1}, r_k, i) \in \mathcal{E}, 1 \le k \le K\}\Big).$$

Graph Encoding with Path-Based Models

- Relation Network (RN)
- KagNet → 直接从图中抽取关系路径，并用序列模型编码的方法

特点：具有可解释性，但扩展性不够好

扩展性不够好的原因：

1) Polynomial （考虑结点数量，图中路径数量的变化呈多项式）
2) Exponential（考虑跳数，图中路径数量的变化呈指数式）

因此，也有一些模型也仅仅使用 One-hop 路径（三元组）来平衡 scalability

Graph Encoding with GNNs

- GNN： $\{\boldsymbol{h}_1, \boldsymbol{h}_2, \ldots, \boldsymbol{h}_n\}$ → a set of node features as input

  $\{\boldsymbol{h}_1', \boldsymbol{h}_2', \ldots, \boldsymbol{h}_n'\}$ → node embeddings via message passing

  $\mathrm{GNN}(\mathcal{G}) = \mathrm{Pool}(\{\boldsymbol{h}_1', \boldsymbol{h}_2', \ldots, \boldsymbol{h}_n'\}).$ → representation for G

- GCN： 为每个结点融合它邻结点的信息
- RGCN： GCN的变体，为每种边的类型定义了特定的权重矩阵$\boldsymbol{W}_r$

$$\boldsymbol{h}_i' = \sigma\left(\left(\sum_{r\in\mathcal{R}} |\mathcal{N}_i^r|\right)^{-1} \sum_{r\in\mathcal{R}} \sum_{j\in\mathcal{N}_i^r} \boldsymbol{W}_r \boldsymbol{h}_j\right),$$

N(i, r) 表示与结点i相连的关系为r的所有邻结点

Graph Encoding with GNNs

- GCNs
- RGCNs

图神经网络

特点：具有可扩展性，但可解释性不够好，缺乏推理的透明度

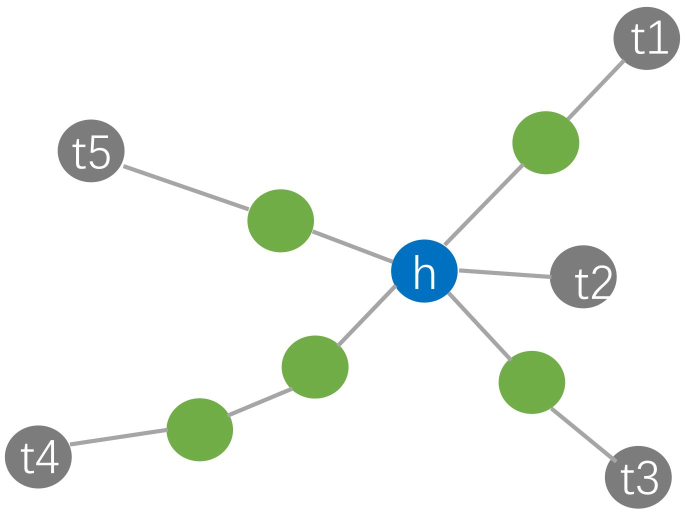
这些模型都没有区分不同邻结点和关系类型的重要性，无法为模型行为解释提供明确的关系路径。

Multi-hop Graph Relation Network (MHGRN)

- ✓ 结合了 Path-based models + GNNs 两者的优点
- ✓ 兼具 可解释性 + 可扩展性

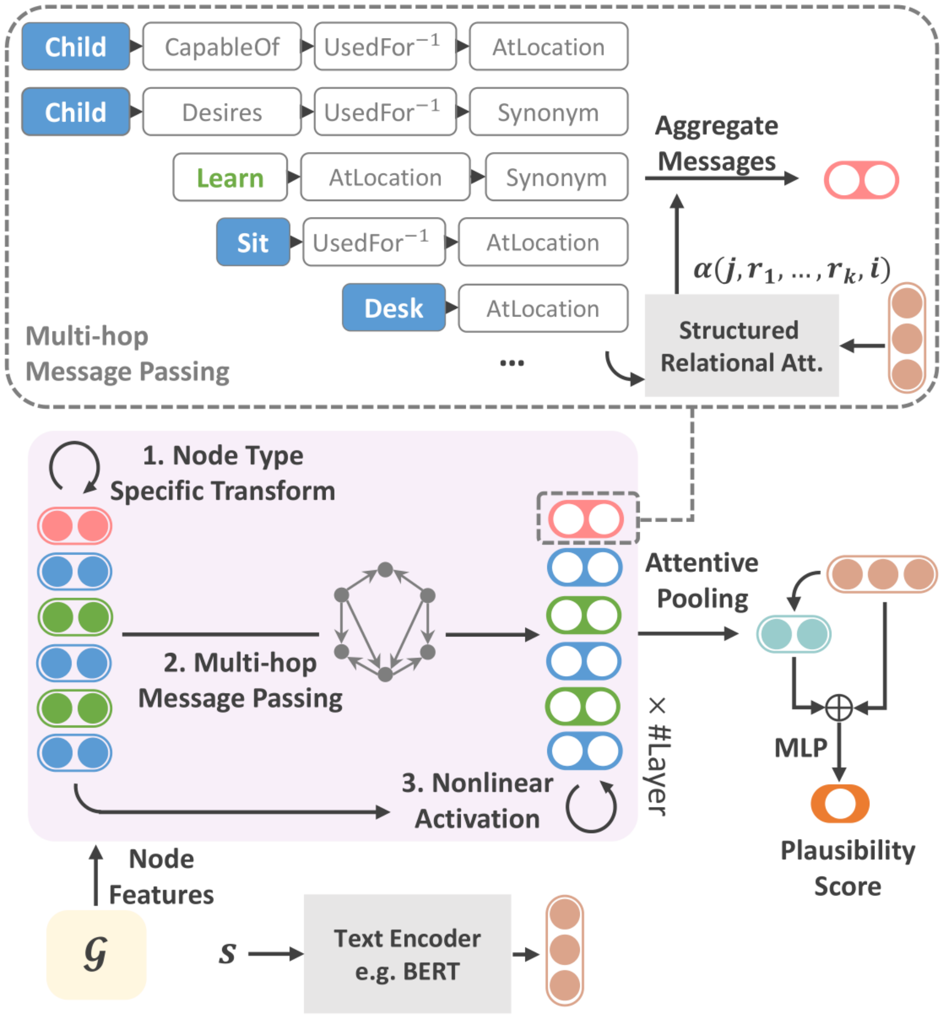
	GCN	RGCN	KagNet	MHGRN
Multi-Relational Encoding	✗	✓	✓	✓
Interpretable	✗	✗	✓	✓
Scalable w.r.t. #node	✓	✓	✗	✓
Scalable w.r.t. #hop	✓	✓	✗	✓

Table 1: **Properties** of our MHGRN and other representative models for graph encoding.



Key Motivation: each node directly **attends** to its multi-hop neighbors

MHGRN: Model Architecture



$$\boldsymbol{x}_i = \boldsymbol{U}_{\phi(i)}\boldsymbol{h}_i + \boldsymbol{b}_{\phi(i)},$$

1. 区分结点类型，做个转换

$$\Phi_k = \{(j, r_1, \dots, r_k, i) \mid (j, r_1, j_1), \cdots, (j_{k-1}, r_k, i) \in \mathcal{E}\} \quad (1 \le k \le K).$$

$$\boldsymbol{z}_i^k = \sum_{(j, r_1, \dots, r_k, i) \in \Phi_k} \alpha(j, r_1, \dots, r_k, i)/d_i^k \cdot \boldsymbol{W}_0^K \cdots \boldsymbol{W}_0^{k+1} \boldsymbol{W}_{r_k}^k \cdots \boldsymbol{W}_{r_1}^1 \boldsymbol{x}_j \quad (1 \le k \le K), \quad (7)$$

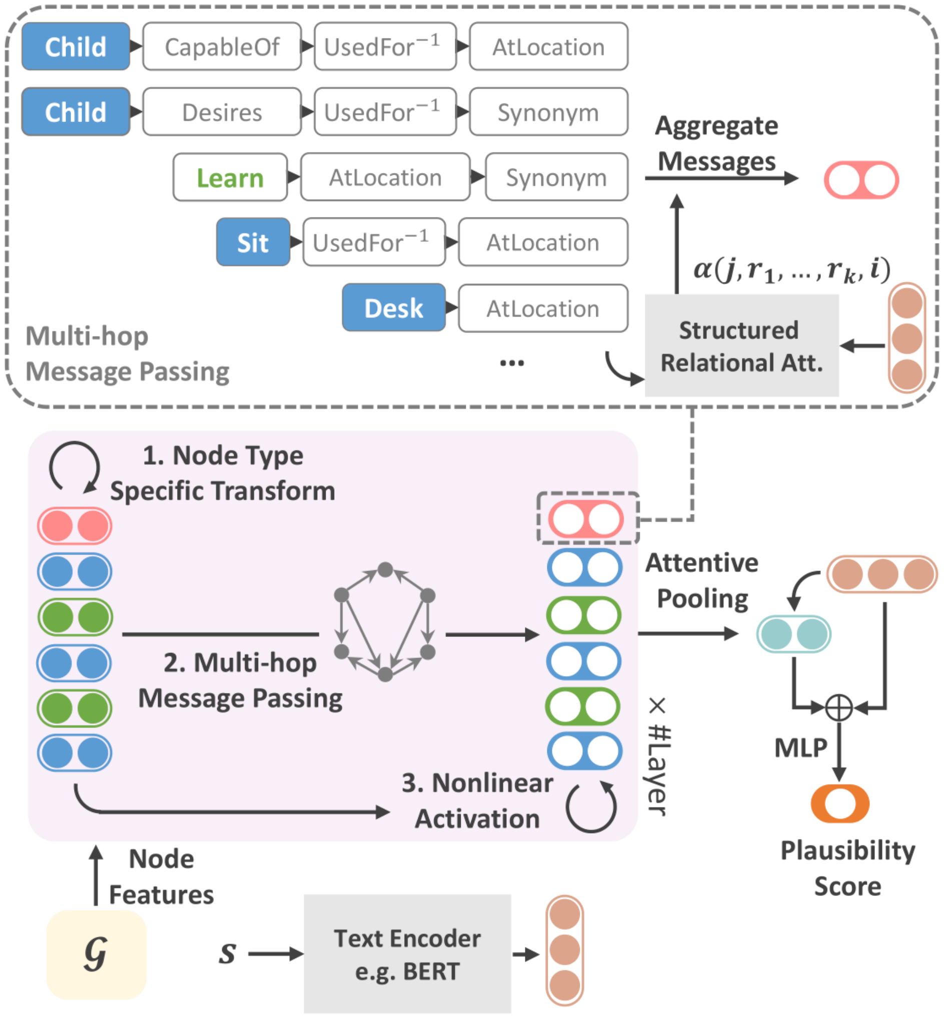
$$\boldsymbol{z}_i = \sum_{k=1}^{K} \mathrm{softmax}(\mathrm{bilinear}(\boldsymbol{s}, \boldsymbol{z}_i^k)) \cdot \boldsymbol{z}_i^k.$$

2. 多跳信息传递

$$\boldsymbol{h}_i' = \sigma\left(\boldsymbol{V}\boldsymbol{h}_i + \boldsymbol{V}'\boldsymbol{z}_i\right),$$

3. 非线性激活函数

MHGRN: Model Architecture



得到G中每个结点的特征表示$\boldsymbol{h}_i'$后，再对所有来自A中的结点$\{\boldsymbol{h}_i' \mid i \in \mathcal{A}\}$ 进行一个pool操作，最终得到图G的表示 $\boldsymbol{g}$

$$\rho(q, a) = \texttt{MLP}(\boldsymbol{s} \oplus \boldsymbol{g})$$

Experiments

- **Datasets**

	Train	Dev	Test
CommonsenseQA (OF)	$9,741$	$1,221$	$1,140$
CommonsenseQA (IH)	$8,500$	$1,221$	$1,241$
OpenbookQA	$4,957$	500	500

Experiments

Methods	BERT-Base		BERT-Large		RoBERTa-Large	
	IHdev-Acc.(%)	IHtest-Acc.(%)	IHdev-Acc.(%)	IHtest-Acc.(%)	IHdev-Acc.(%)	IHtest-Acc.(%)
w/o KG	57.31 (±1.07)	53.47 (±0.87)	61.06 (±0.85)	55.39 (±0.40)	73.07 (±0.45)	68.69(±0.56)
RGCN (Schlichtkrull et al., 2018)	56.94 (±0.38)	54.50 (±0.56)	62.98 (±0.82)	57.13 (±0.36)	72.69 (±0.19)	68.41 (±0.66)
GconAttn (Wang et al., 2019)	57.27 (±0.70)	54.84 (±0.88)	63.17 (±0.18)	57.36 (±0.90)	72.61(±0.39)	68.59 (±0.96)
KagNet† (Lin et al., 2019)	55.57	56.19	62.35	57.16	-	-
RN (1-hop)	58.27 (±0.22)	56.20 (±0.45)	63.04 (±0.58)	58.46 (±0.71)	74.57 (±0.91)	69.08 (±0.21)
RN (2-hop)	59.81 (±0.76)	56.61 (±0.68)	63.36 (±0.26)	58.92 (±0.14)	73.65 (±3.09)	69.59 (±3.80)
MHGRN	60.36 (±0.23)	**57.23** (±0.82)	63.29(±0.51)	**60.59** (±0.58)	74.45 (±0.10)	**71.11** (±0.81)

Performance comparison on **CommonsenseQA (IH)**

Experiments

Methods	Single	Ensemble
UnifiedQA† (Khashabi et al., 2020)	**79.1**	-
RoBERTa†	72.1	72.5
RoBERTa + KEDGN†	72.5	74.4
RoBERTa + KE†	73.3	-
RoBERTa + HyKAS 2.0† (Ma et al., 2019)	73.2	-
RoBERTa + FreeLB† (Zhu et al., 2020)	72.2	73.1
XLNet + DREAM†	66.9	73.3
XLNet + GR† (Lv et al., 2019)	75.3	-
ALBERT† (Lan et al., 2019)	-	**76.5**
RoBERTa + MHGRN ($K = 2$)	75.4	**76.5**

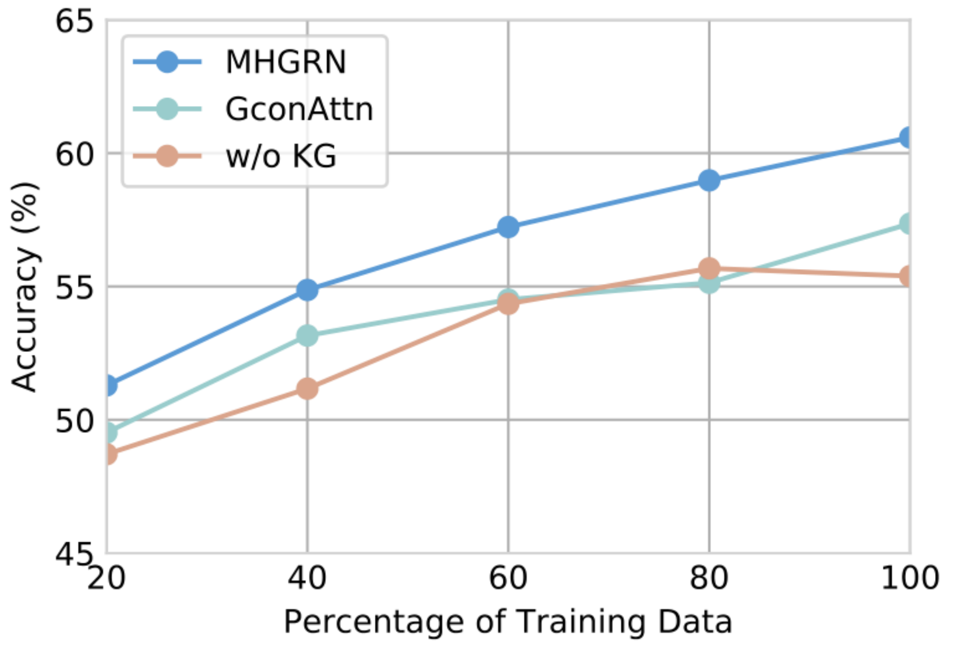
Performance comparison on **CommonsenseQA (OF)**

Experiments

Methods	Dev (%)	Test (%)
T5-3B† (Raffel et al., 2019)	-	83.20
UnifiedQA† (Khashabi et al., 2020)	-	**87.20**
RoBERTa-Large (w/o KG)	66.76 (±1.14)	64.80 (±2.37)
+ RGCN	64.65 (±1.96)	62.45 (±1.57)
+ GconAttn	64.30 (±0.99)	61.90 (±2.44)
+ RN (1-hop)	64.85 (±1.11)	63.65 (±2.31)
+ RN (2-hop)	67.00 (±0.71)	65.20 (±1.18)
+ MHGRN (K = 3)	68.10 (±1.02)	**66.85** (±1.19)
AristoRoBERTaV7†	79.2	77.8
+ MHGRN (K = 3)	78.6	**80.6**

Performance comparison on **OpenbookQA**

Experiments



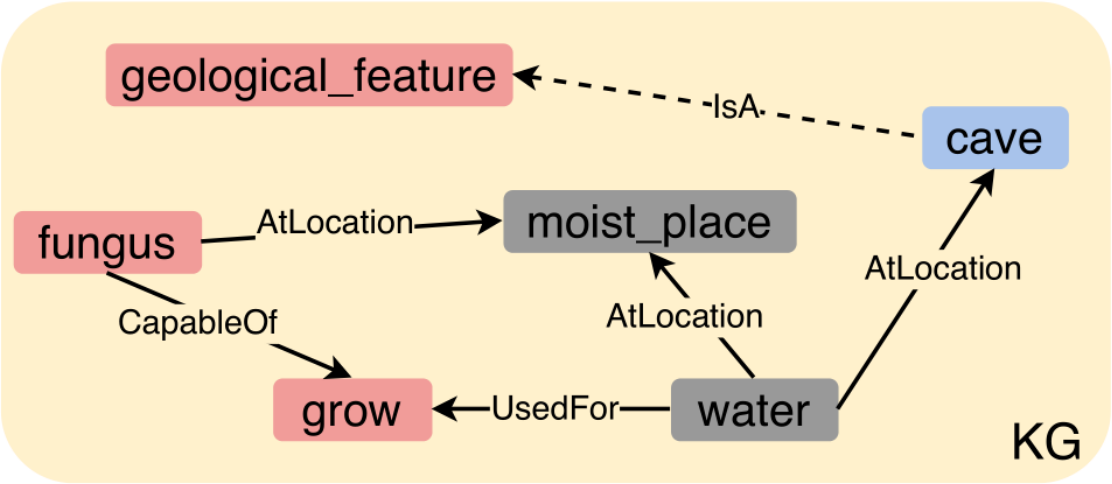
Impact of the Amount of Training Data

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

Peifeng Wang[1,3], **Nanyun Peng**[1,2,3], **Filip Ilievski**[3], **Pedro Szekely**[1,3], **Xiang Ren**[1,3]
[1]Department of Computer Science, University of Southern California
[2]Department of Computer Science, University of California, Los Angeles
[3]Information Sciences Institute, University of Southern California
{peifengw,xiangren}@usc.edu, violetpeng@cs.ucla.edu
{ilievski,pszekely}@isi.edu

EMNLP 2020
https://www.aclweb.org/anthology/2020.findings-emnlp.369.pdf

问题描述



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

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

The missing link:

(cave, IsA, geological feature)

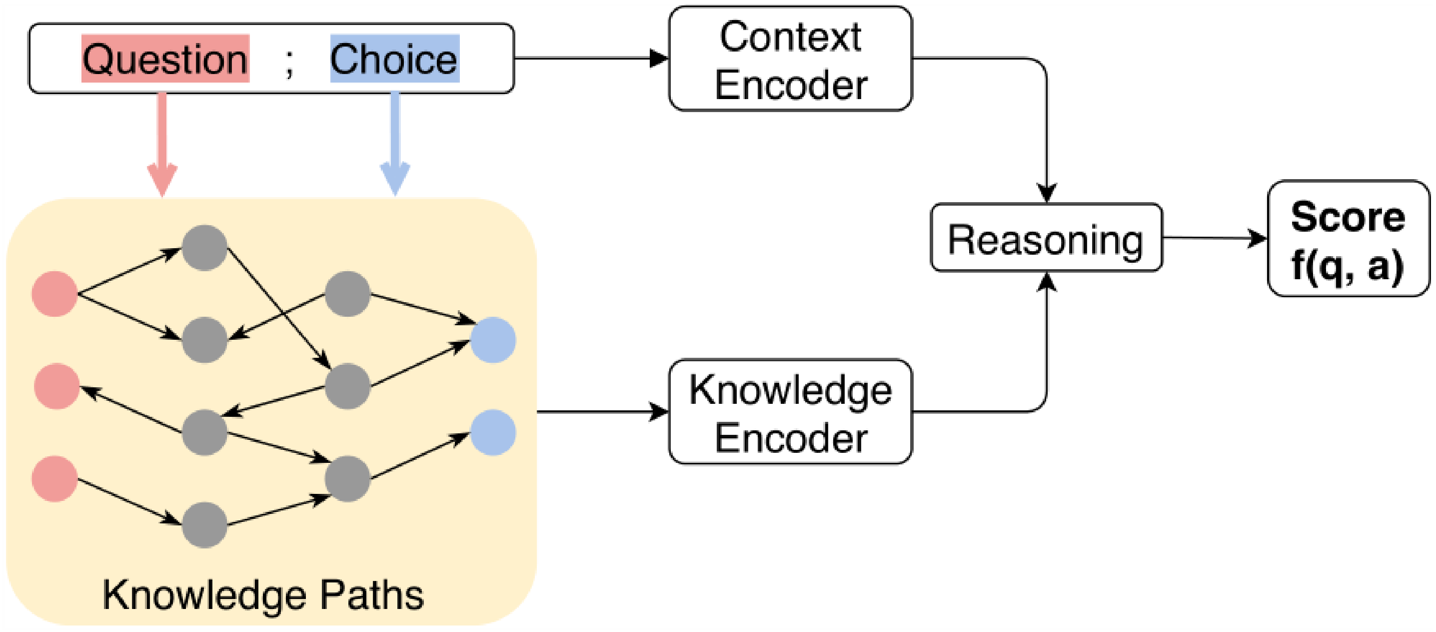
问题背景

✓ **Existing systems retrieve knowledge from a KG，the challenges:**

1) Sparsity

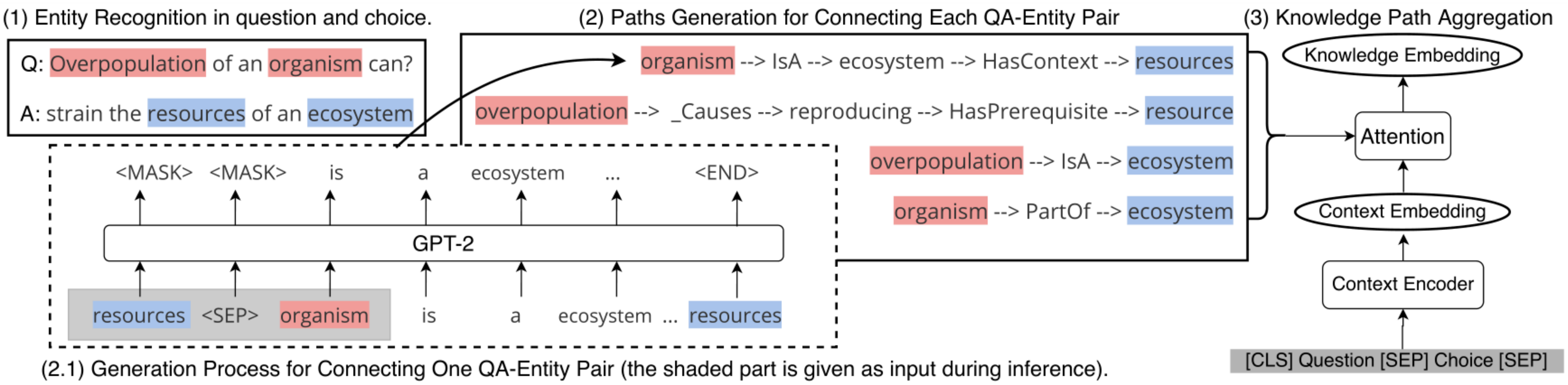
2) Noisy

KG-augmented QA Framework



关系图，本文不从静态知识图谱（eg. ConceptNet）中抽取得到，改用GPT-2动态生成

Model Architecture



怎么生成和任务相关的
知识路径呢？

Knowledge Path Sampling（知识路径采样）

从已有的静态图谱（KG）中进行路径采样，用来微调GPT-2

为了保证采样路径的质量，制定了两种策略

✓ **Relevance (相关性)**

Define useful relation types，filter out the remaining ones

✓ **Informativeness (信息性)**

All relation types in a path to be distinct

Knowledge Path Sampling（知识路径采样）

使用了两种采样方法

✓ **Local sampling (局部采样)**

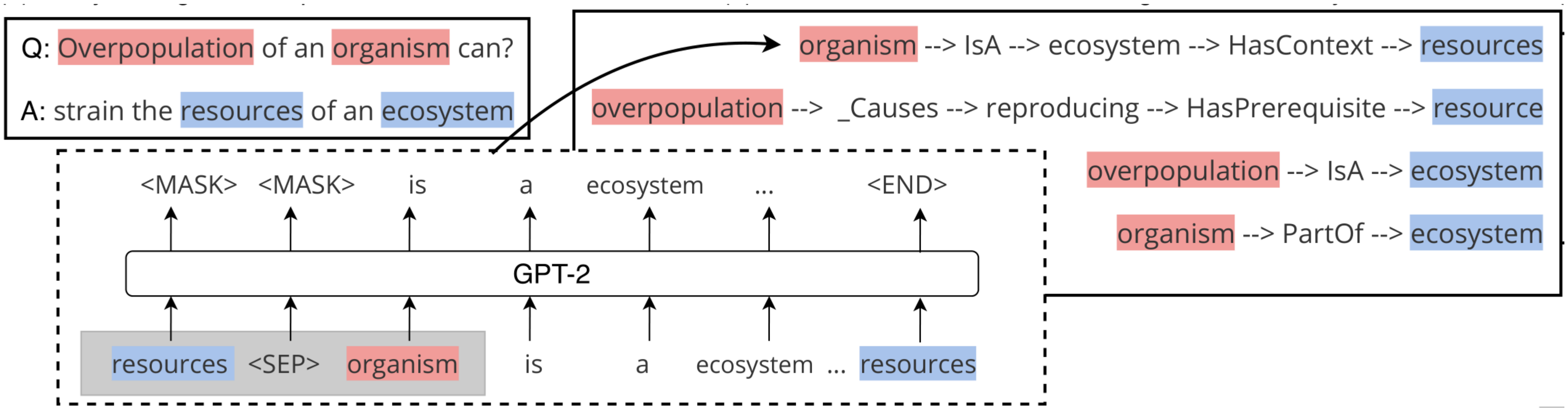
Path的起始结点是任务训练集中Q和A中的实体，并从它们开始进行随机游走，得到的路径

✓ **Global sampling (全局采样)**

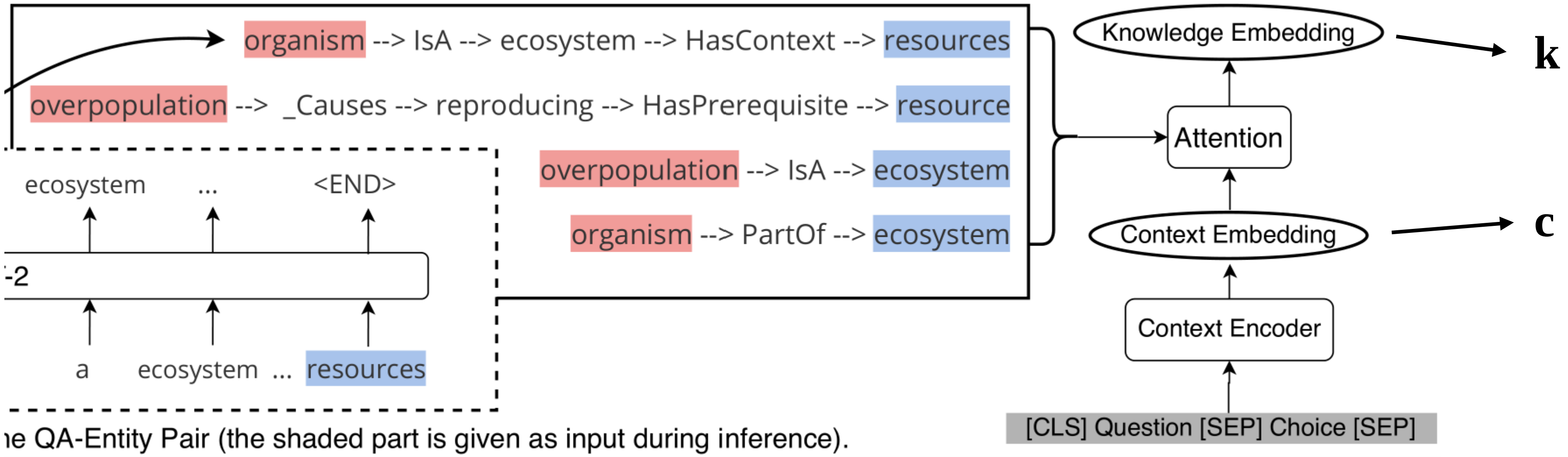
Path的起始结点是随机采样的一些实体，并从它们开始进行随机游走，得到一些局部KG以外的路径，用于生成器的泛化

基于GPT-2的路径生成器的构建

用采样的路径上对GPT-2进行微调，之后便可以用来生成我们任务数据集相关的知识路径



Reasoning Module



Q和A的匹配得分： $f(q, a) = \mathbf{W}_{cls} \cdot [\mathbf{c}; \mathbf{k}] + \mathbf{b}_{cls},$

Experiments

- **Datasets**

	Train	Dev	Test
CommonsenseQA (OF)	$9,741$	$1,221$	$1,140$
CommonsenseQA (IH)	$8,500$	$1,221$	$1,241$
OpenbookQA	$4,957$	500	500

Experiments

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)

Test accuracy with varying proportions of **CommonsenseQA (IH)**

Experiments

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	<u>78.2</u>

Test accuracy on **CommonsenseQA (OF)**

Experiments

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)

Test accuracy on **OpenBookQA**

THE END

2020.12.03