Diversifying Content Generation for Commonsense Reasoning with Mixture of Knowledge Graph Experts

Wenhao Yu♣, Chenguang Zhu♠, Lianhui Qin♥,
Zhihan Zhang♣, Tong Zhao♣, Meng Jiang♣
♣University of Notre Dame ♥University of Washington
♠Microsoft Cognitive Services Research
♣{wyu1, zzhang23, tzhao2, mjiang2}@nd.edu
♠chezhu@microsoft.com
♥lianhuiq@cs.washington.edu

ACL2022

Code:https://github.com/DM2-ND/MoKGE

2022. 5. 18 • ChongQing

Leibniz-Institut für Sozialwissenschaften

Reported by Yang Peng

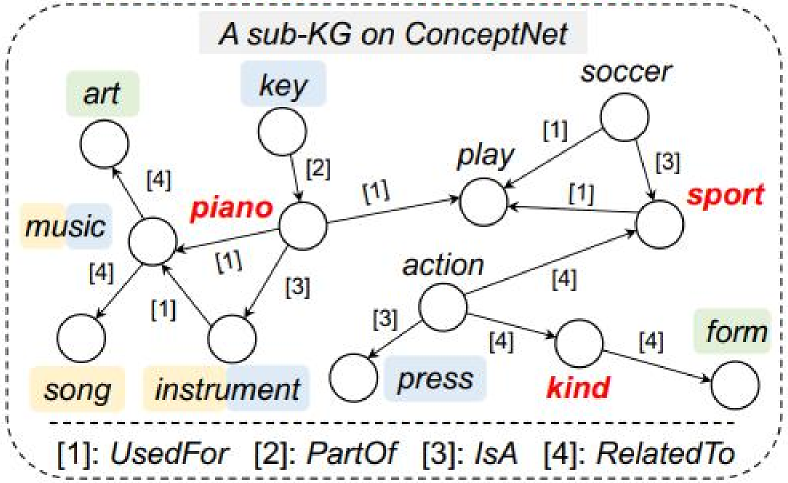
1.Introduction

2.Method

3.Experiments

Introduction

Input: Piano is a kind of sport .



Outputs: *3* different explanations

(1) You can produce music when pressing keys on the piano, so it is an instrument .
(2) Piano is a musical instrument used in songs to produce different musical tones .
(3) Piano is a kind of art form .

Figure 1: An example of diverse commonsense explanation generation. It aims at generating multiple reasonable explanations given a counterfactual statement. Relevant concepts on the commonsense KG (in shade) can help to perform diverse knowledge reasoning.

Method

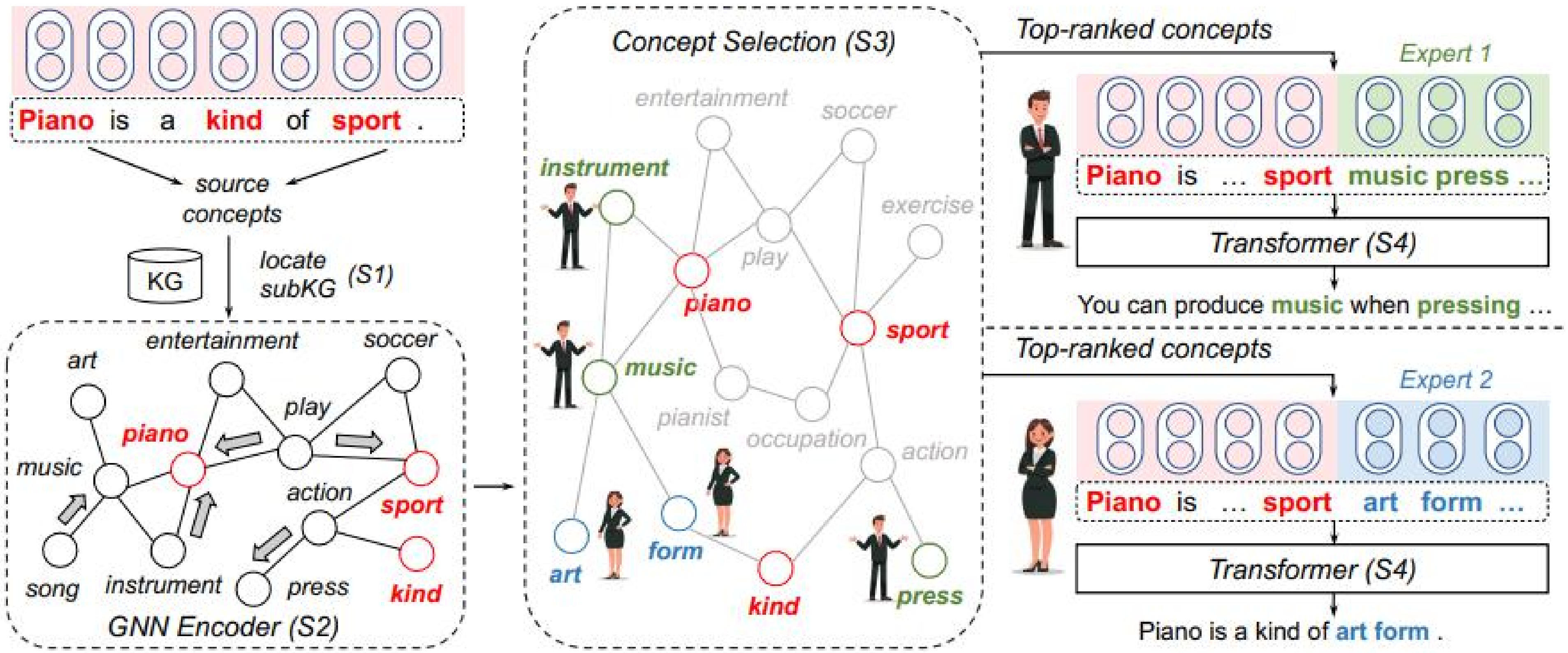
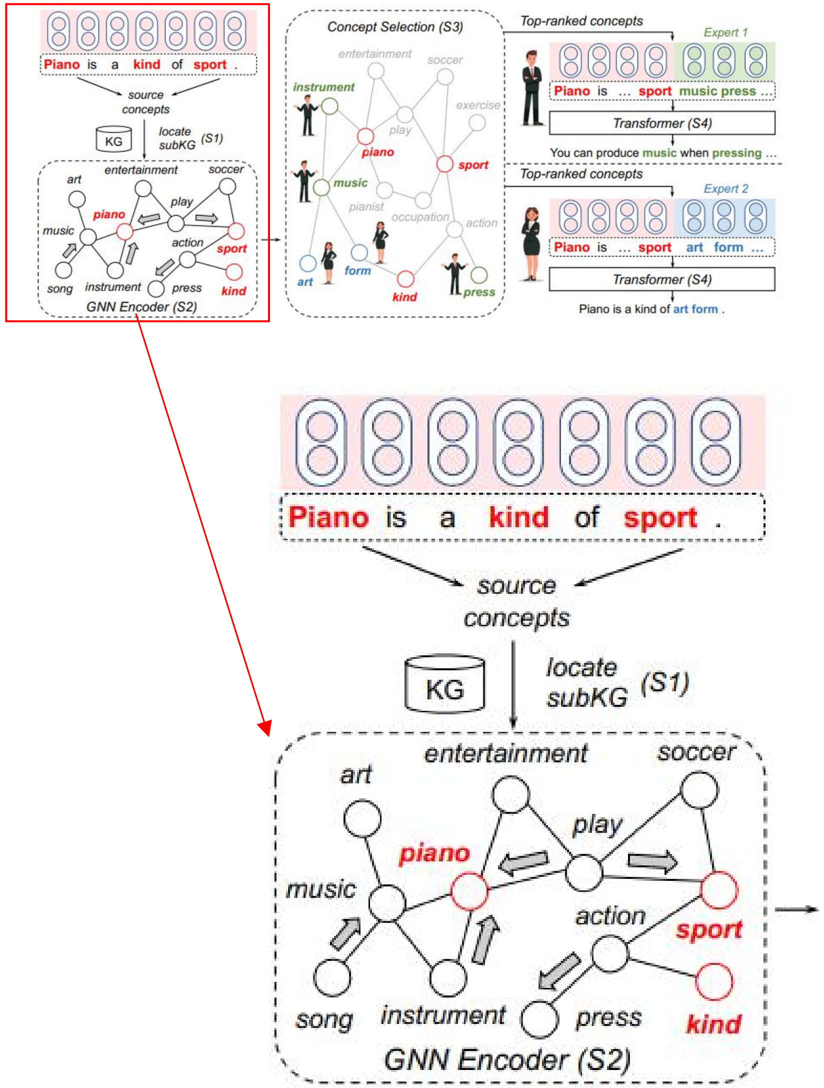


Figure 2: The overall architecture of MoKGE. The MoKGE consists of four steps: (S1) the model constructs a sequence-associated subgraph from the commonsense KG; (S2) a relational-GCN iteratively updates the representation of a concept node by aggregating information from its neighboring nodes and edges; (S3) each knowledge expert selects different salient concepts that should be considered during generation; (S4) the model generates the outputs by integrating the token embeddings of the input sequence and the top-ranked entities.



Method

Sequence-aware subgraph construction

commonsense knowledge graph $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$

a sequence-associated subgraph $\mathcal{G}_x = \{\mathcal{V}_x, \mathcal{E}_x\}$

Multi-relational graph encoding

node embedding updated

$$\mathbf{o}_v^l = \frac{1}{|\mathcal{N}(v)|} \sum_{(u,v,r)\in\mathcal{E}} \mathbf{W}_N^l \phi(\mathbf{h}_u^l, \mathbf{h}_r^l), \quad (1)$$

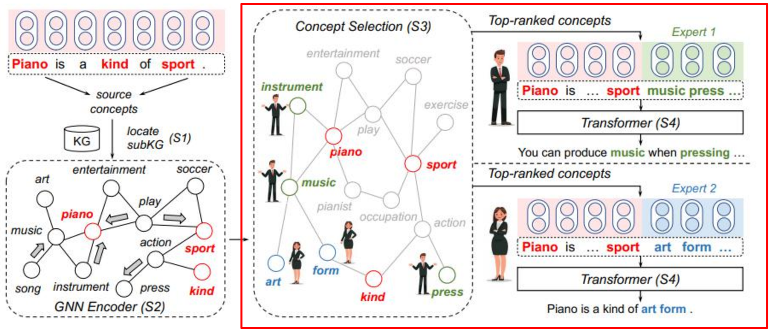
$$\mathbf{h}_v^{l+1} = \mathrm{ReLU}(\mathbf{o}_v^l + \mathbf{W}_S^l \mathbf{h}_v^l), \quad (2)$$

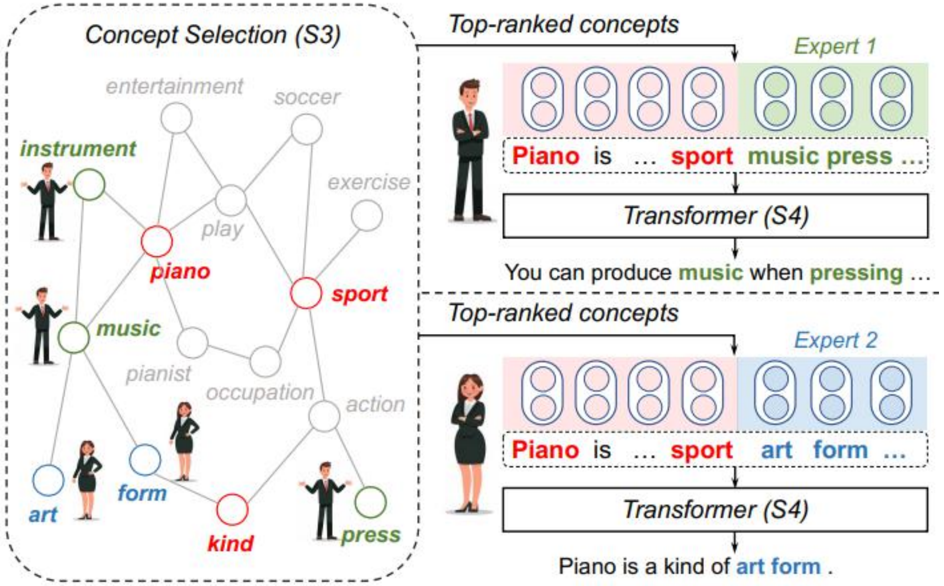
$$\phi(\mathbf{h}_u, \mathbf{h}_r) = \mathbf{h}_u - \mathbf{h}_r$$

relation embedding updated

$$\mathbf{h}_r^{l+1} = \mathbf{W}_R^l \mathbf{h}_r^l. \quad (3)$$

Finally, we obtain concept embedding $\mathbf{h}_v^L$ that encodes the sequence-associated subgraph context.





Method

Concept selection on knowledge graph

$$p_v = Pr[v \text{ is selected} | x] = \mathrm{MLP}(\mathbf{h}_v^L).$$

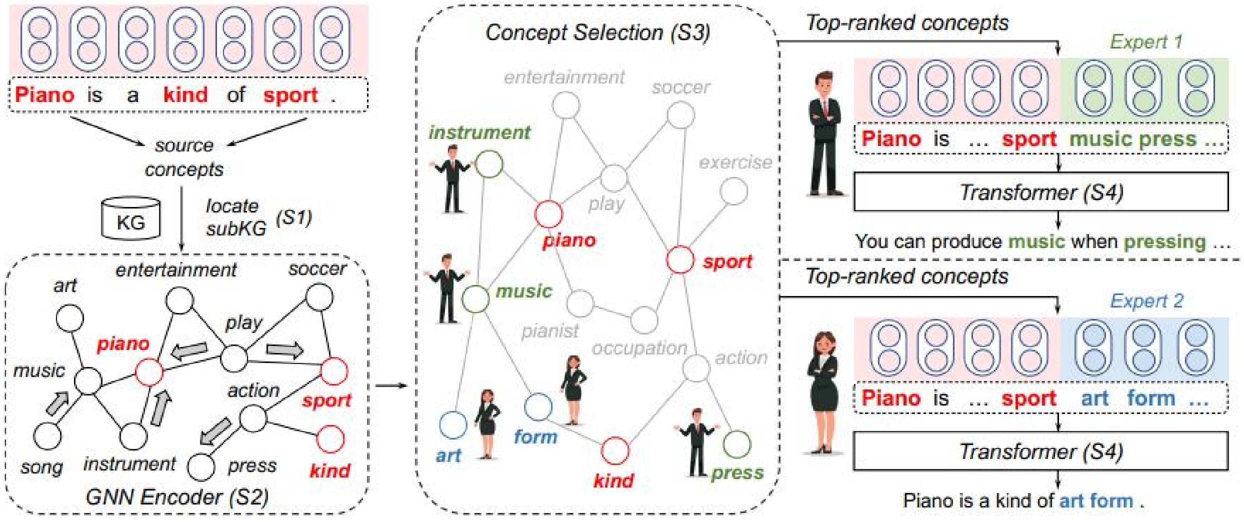
$$\mathcal{L}_{\text{concept}} = -\Big(\sum_{v \in \mathcal{V}_x \cap C_y} v \log p_v + \sum_{v \in \mathcal{V}_x - C_y} (1-v) \log(1-p_v) \Big). \quad (4)$$

top-N ranked concepts on the subgraph G_x (denoted as $v_1, ..., v_N$)

$$\mathcal{L}_{\text{generation}} = -\log p(y | x, v_1, \cdots, v_N) = -\sum_{t=1}^{|y|} \log p(y_t | x, v_1, \cdots, v_N, y_{<t}). \quad (5)$$

$$\mathcal{L} = \mathcal{L}_{\text{generation}} + \lambda \cdot \mathcal{L}_{\text{concept}}. \quad (6)$$

Method



MoE-Promoted Diverse Generation

MoE module introduces a multinomial latent variable

$$z \in \{1, \cdots, K\}$$

$$p(y|x, \mathcal{G}_x) = \sum_{z=1}^{K} p(z|x, \mathcal{G}_x) p(y|z, x, \mathcal{G}_x). \quad (7)$$

Training. We minimize the loss function (in Eq.(6)) using the MoE decomposition,

$$\nabla \log p(y|x, \mathcal{G}_x) \quad (8)$$

$$= \sum_{z=1}^{K} p(z|x, y, \mathcal{G}_x) \cdot \nabla \log p(y, z|x, \mathcal{G}_x),$$

Experiments

Table 2: Diversity and quality evaluation on the **ComVE** (upper part) and α-**NLG** (lower part) datasets. Each model is required to generate three outputs. All experiments are run three times with different random seeds, and the average results on the test set is calculated as the final performance, with standard deviations as subscripts.

Methods	Model Variant	Concept diversity		Pairwise diversity		Corpus diversity		Quality	
		#Uni.C($\Uparrow$)	Jaccard ($\Downarrow$)	SB-3 ($\Downarrow$)	SB-4 ($\Downarrow$)	D-2($\Uparrow$)	E-4($\Uparrow$)	B-4 ($\Uparrow$)	R-L ($\Uparrow$)
CVAE	z = 16	$4.56_{0.1}$	$64.74_{0.3}$	$66.66_{0.4}$	$62.83_{0.5}$	$33.75_{0.5}$	$9.13_{0.1}$	$16.67_{0.3}$	$41.52_{0.3}$
	z = 32	$5.03_{0.3}$	$47.27_{0.8}$	$59.20_{1.3}$	$54.30_{1.5}$	$32.86_{1.1}$	$9.07_{0.5}$	$17.04_{0.2}$	$42.17_{0.5}$
	z = 64	$4.67_{0.0}$	$54.69_{0.8}$	$55.02_{0.8}$	$49.58_{1.0}$	$32.55_{0.5}$	$9.07_{0.2}$	$15.54_{0.4}$	$41.03_{0.3}$
Truncated sampling	k = 5	$4.37_{0.0}$	$71.38_{0.7}$	$74.20_{0.2}$	$71.38_{0.2}$	$31.32_{0.4}$	$9.18_{0.1}$	$16.44_{0.2}$	$40.99_{0.2}$
	k = 20	$4.60_{0.0}$	$63.42_{1.2}$	$64.47_{2.1}$	$60.33_{2.4}$	$33.69_{0.6}$	$9.26_{0.1}$	$17.70_{0.2}$	$42.58_{0.5}$
	k = 50	$4.68_{0.1}$	$60.98_{1.8}$	$61.39_{2.4}$	$56.93_{2.8}$	$34.80_{0.3}$	$9.29_{0.1}$	$17.48_{0.4}$	$42.44_{0.5}$
Nucleus sampling	p = .5	$4.19_{0.1}$	$72.78_{1.0}$	$77.66_{0.8}$	$75.14_{0.9}$	$28.36_{0.6}$	$9.05_{0.3}$	$16.09_{0.6}$	$40.95_{0.5}$
	p = .75	$4.41_{0.1}$	$67.01_{1.7}$	$71.41_{2.5}$	$68.22_{2.9}$	$31.21_{0.3}$	$9.16_{0.1}$	$17.07_{0.5}$	$41.88_{0.7}$
	p = .95	$4.70_{0.1}$	$61.92_{2.6}$	$63.43_{3.4}$	$59.23_{3.8}$	$34.17_{0.3}$	$9.27_{0.2}$	$17.68_{0.4}$	$42.60_{0.8}$
MoE	embed	$5.41_{0.0}$	$\underline{47.55}_{0.5}$	$33.64_{0.2}$	$\underline{28.21}_{0.1}$	$46.57_{0.2}$	$9.61_{0.1}$	$18.66_{0.5}$	$\underline{43.72}_{0.2}$
	prompt	$\underline{5.45}_{0.2}$	$47.54_{0.4}$	$\underline{33.42}_{0.3}$	$28.40_{0.3}$	$46.93_{0.2}$	$9.60_{0.2}$	$18.91_{0.4}$	$43.71_{0.5}$
MoKGE (ours)	embed	$5.35_{0.2}$	$48.18_{0.5}$	$35.36_{1.1}$	$29.71_{1.2}$	$\underline{47.51}_{0.4}$	$\underline{9.63}_{0.1}$	$\mathbf{19.13}_{0.1}$	$43.70_{0.1}$
	prompt	$\mathbf{5.48}_{0.2}$	$\mathbf{44.37}_{0.4}$	$\mathbf{30.93}_{0.9}$	$\mathbf{25.30}_{1.1}$	$\mathbf{48.44}_{0.2}$	$\mathbf{9.67}_{0.2}$	$\underline{19.01}_{0.1}$	$\mathbf{43.83}_{0.3}$
Human		$6.27_{0.0}$	$26.49_{0.0}$	$12.36_{0.0}$	$8.01_{0.0}$	$63.02_{0.0}$	$9.55_{0.0}$	$100.0_{0.0}$	$100.0_{0.0}$

* Metrics: SB-3/4: Self-BLEU-3/4 ($\Downarrow$), D-2: Distinct-2 ($\Uparrow$), E-4: Entropy-4 ($\Uparrow$), B-4: BLEU-4 ($\Uparrow$), R-L: ROUGE-L ($\Uparrow$)

Experiments

		#Uni.C(⇑)	Jaccard (⇓)	SB-3 (⇓)	SB-4 (⇓)	D-2(⇑)	E-4(⇑)	B-4 (⇑)	R-L (⇑)
CVAE	z = 16	$4.80_{0.0}$	$56.88_{0.1}$	$67.89_{0.4}$	$64.72_{0.5}$	$26.27_{0.2}$	$10.34_{0.0}$	$13.64_{0.1}$	$37.96_{0.1}$
	z = 32	$5.05_{0.0}$	$50.92_{0.4}$	$62.08_{0.2}$	$58.25_{0.3}$	$26.67_{0.1}$	$10.36_{0.0}$	$13.35_{0.1}$	$37.73_{0.1}$
	z = 64	$5.14_{0.0}$	$47.04_{0.7}$	$57.87_{0.4}$	$53.61_{0.4}$	$24.91_{0.1}$	$10.21_{0.1}$	$11.77_{0.1}$	$36.35_{0.2}$
Truncated sampling	k= 5	$4.86_{0.1}$	$72.78_{1.1}$	$67.09_{1.0}$	$63.82_{1.1}$	$25.47_{0.3}$	$10.44_{0.1}$	$13.33_{0.2}$	$38.07_{0.2}$
	k= 20	$5.48_{0.1}$	$45.65_{1.8}$	$54.65_{2.1}$	$50.36_{2.4}$	$29.30_{0.5}$	$10.62_{0.2}$	$14.12_{0.7}$	$38.76_{0.6}$
	k= 50	$5.53_{0.0}$	$45.84_{0.5}$	$52.11_{3.7}$	$47.75_{4.2}$	$30.08_{0.3}$	$10.64_{0.1}$	$14.01_{0.8}$	$\mathbf{38.98}_{0.6}$
Nucleus sampling	p= .5	$4.19_{0.1}$	$62.54_{1.8}$	$73.34_{0.3}$	$71.01_{0.3}$	$25.49_{0.0}$	$10.46_{0.0}$	$11.71_{0.1}$	$36.53_{0.2}$
	p= .75	$5.13_{0.0}$	$54.25_{0.6}$	$64.49_{0.4}$	$61.45_{0.5}$	$27.72_{0.1}$	$10.54_{0.1}$	$12.63_{0.0}$	$37.48_{0.1}$
	p= .95	$5.49_{0.0}$	$46.76_{0.5}$	$56.32_{0.5}$	$52.44_{0.6}$	$29.92_{0.1}$	$10.63_{0.0}$	$13.53_{0.2}$	$38.42_{0.3}$
MoE	embed	$6.22_{0.1}$	$\underline{29.18}_{0.4}$	$29.02_{1.0}$	$24.19_{1.0}$	$36.22_{0.3}$	$10.84_{0.0}$	$\mathbf{14.31}_{0.2}$	$\underline{38.91}_{0.2}$
	prompt	$6.05_{0.1}$	$29.34_{1.2}$	$\underline{28.05}_{2.0}$	$\underline{23.18}_{1.9}$	$36.71_{0.1}$	$10.85_{0.0}$	$\underline{14.26}_{0.3}$	$38.78_{0.4}$
MoKGE (ours)	embed	$\underline{6.27}_{0.2}$	$30.46_{0.8}$	$29.17_{1.5}$	$24.04_{1.6}$	$\mathbf{38.15}_{0.3}$	$\mathbf{10.90}_{0.1}$	$13.74_{0.2}$	$38.06_{0.2}$
	prompt	$\mathbf{6.35}_{0.1}$	$\mathbf{28.06}_{0.6}$	$\mathbf{27.40}_{2.0}$	$\mathbf{22.43}_{2.4}$	$\underline{38.01}_{0.6}$	$\underline{10.88}_{0.2}$	$14.17_{0.2}$	$38.82_{0.7}$
Human		$6.62_{0.0}$	$12.43_{0.0}$	$10.36_{0.0}$	$6.04_{0.0}$	$53.57_{0.0}$	$10.84_{0.0}$	$100.0_{0.0}$	$100.0_{0.0}$

* Metrics: SB-3/4: Self-BLEU-3/4 (⇓), D-2: Distinct-2 (⇑), E-4: Entropy-4 (⇑), B-4: BLEU-4 (⇑), R-L: ROUGE-L (⇑)

Experiments

Table 3: Ablation studies. When not suing MoE (line –w/o MoE), we set beam as three to generate three outputs.

Methods	ComVE (left part: diversity; right part: quality)					α-NLG (left part: diversity; right part: quality)				
	SB-4 (⇓)	D-2 (⇑)	E-4 (⇑)	B-4 (⇑)	R-L (⇑)	SB-4 (⇓)	D-2 (⇑)	E-4 (⇑)	B-4 (⇑)	R-L (⇑)
MoKGE	$\mathbf{25.30}_{1.1}$	$\mathbf{48.44}_{0.2}$	$\mathbf{9.67}_{0.2}$	$\mathbf{19.01}_{0.1}$	$\mathbf{43.83}_{0.3}$	$\mathbf{22.43}_{2.4}$	$\mathbf{38.01}_{0.6}$	$\mathbf{10.88}_{0.2}$	$14.17_{0.2}$	$\mathbf{38.82}_{0.7}$
⊢ w/o KG	$28.40_{0.3}$	$46.93_{0.2}$	$9.60_{0.2}$	$18.91_{0.4}$	$43.71_{0.5}$	$23.18_{1.9}$	$36.71_{0.1}$	$10.85_{0.0}$	$\mathbf{14.26}_{0.3}$	$38.78_{0.4}$
⊢ w/o MoE	$74.15_{0.2}$	$31.92_{0.1}$	$9.14_{0.0}$	$15.87_{0.1}$	$40.24_{0.2}$	$77.34_{0.2}$	$19.19_{0.1}$	$10.10_{0.0}$	$12.84_{0.1}$	$37.52_{0.2}$

Experiments

Table 4: Human evaluations by independent scoring based on *diveristy*, *quality*, *flency* and *grammar*. In addition, * indicates p-value < 0.05 under paired t-test between MoKGE and baseline methods.

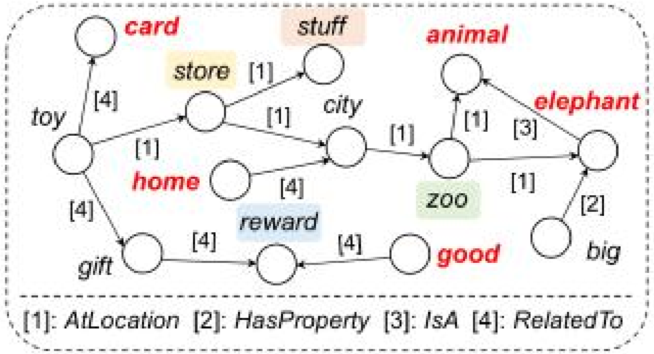
Methods	ComVE			α-NLG		
	Diversity	Quality	Flu. & Gra.	Diversity	Quality	Flu. & Gra.
Truncated samp.	2.15±0.76	2.22±1.01	3.47±0.75	2.31±0.76	2.63±0.77	3.89±0.36
Nucleus samp.	2.03±0.73	**2.29**±1.03	**3.52**±0.70	2.39±0.73	**2.67**±0.72	**3.91**±0.28
MoKGE (ours)	**2.63**±0.51*	2.10±0.99	3.46±0.81	**2.66**±0.51*	2.57±0.71	3.87±0.34
Human Ref.	2.60±0.59	3.00	4.00	2.71±0.57	3.00	4.00

Table 5: Human evaluations by pairwise comparison: MoKGE v.s. two baseline methods based on *diversity*.

Against methods	ComVE			α-NLG		
	Win (%)	Tie (%)	Lose (%)	Win (%)	Tie (%)	Lose (%)
v.s. Truncated samp.	**47.85**±5.94	37.09±4.56	15.06±3.31	**45.35**±5.06	43.19±2.78	11.46±2.31
v.s. Nucleus samp.	**54.30**±4.62	36.02±2.74	9.68±3.48	41.53±1.55	**46.99**±2.04	11.48±2.36

Experiments

α-NLG -- Input: Billy had received good grades on his report card. [?]. He decided as he got home that elephants were his new favorite animal.



Nucleus sampling

(1) Billy wanted to go to the zoo and see elephants.
(2) Billy was excited to go on his trip to the zoo.
(3) Billy went to the zoo to see the animals.

MoE (Shen et al.,)

(1) Billy went to the zoo to see the animals.
(2) Billy was excited to go to the zoo with his friends.
(3) Billy's parents took him to the zoo to see elephants.

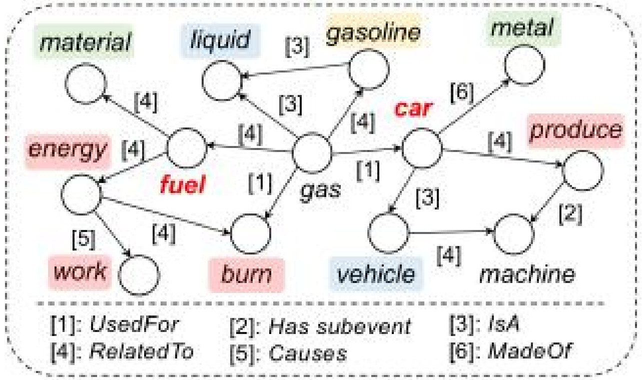
MoKGE (ours)

(1) Billy's parents took him to the zoo as a reward.
(2) Billy wanted to go to the zoo. He saw elephants.
(3) Billy went to the store and bought an elephant.

Human references

(1) Billy's parents sent him on an African safari for a reward.
(2) He went to the zoo later in the day and saw elephants.
(3) His mother stopped by the store and bought him a stuffed elephant.

ComVE -- Input: *Cars are made of fuel.* Goal (explanation for sense-making): [?].



Nucleus sampling

(1) Cars are made of rubber. Fuel is not used to make cars.
(2) Cars are made of aluminum, which is not fuel.
(3) Cars are powered by electric motors and not by fuel.

MoE (Shen et al.,)

(1) Cars are made of metal. but not fuel.
(2) Cars are made of aluminum, not made by fuel.
(3) Fuel is used to make cars more efficient, not less so.

MoKGE (ours)

(1) Fuel is not a vehicle material.
(2) Fuel is not used to make cars. They use gasoline.
(3) Cars are not made of fuel. They are made of metal.

Human references

(1) Cars are not made of fuel.
(2) Cars burn fuel to produce energy and work.
(3) Fuel is a liquid which cannot make cars.

Figure 3: Case studies. MoKGE can produce diverse knowledge reasoning on commonsense KG, select different relevant concepts (in shades of different colors), then generate diverse outputs. The outputs diversity of MoKGE is significantly better than that of beam search and nucleus sampling, and close to human performance.

Thank you!