中北大学学报(自然科学版)
JOURNAL OF NORTH UNIVERSITY OF CHINA(NATURAL SCIENCE EDITION)
2026年 第47卷 第1期(总第225期)
Vol. 47 No. 1 2026(Sum No. 225)

文章编号：1673-3193(2026)01-0062-09 J. North Univ. China, Nat. Sci. Ed., 2026, 47(1): 62-70.

基于多粒度知识的无监督常识问答

杨陟卓，王年楷

（山西大学 计算机与信息技术学院，山西 太原 030006）

摘　要：常识性问答(Commonsense Question Answering，CQA)是一项比传统问答任务更具挑战性的自然语言理解任务，它要求模型具备更强的常识推理能力。目前，基于无监督方法的常识问答在若干数据集上取得了较好的性能，但这些方法难以充分挖掘和利用常识知识，限制了模型在复杂场景下的推理能力。针对这一问题，本文提出了一种新颖的无监督常识问答方法，其核心优势在于通过无监督学习有效整合外部常识知识，从而提升模型的泛化能力和推理深度。首先，该方法对问题进行分类，区分科学常识问题与日常事件问题；随后，根据问题类型生成相应的知识前缀；接着，将知识前缀输入预训练语言模型，通过大模型提示生成多粒度的常识知识；最后，利用多粒度知识辅助问答推理模块进行答案生成。采用无监督方法不仅可以减少对标注数据的依赖，还能更好地适应多样化的常识场景，体现了其在实际应用中的灵活性和普适性。实验结果表明，所提方法在相关数据集上显著优于基线模型，验证了其在无监督常识问答任务中的正确性和合理性。

关键词：常识问答；大模型提示；知识生成；答案推理

中图分类号：TP391.41　　**文献标识码：**A　　**doi:** 10.62756/jnuc.issn.1673-3193.2025.03.0013

引用格式：杨陟卓，王年楷. 基于多粒度知识的无监督常识问答[J]. 中北大学学报(自然科学版)，2026，47(1)：62-70.

YANG Zhizhuo，WANG Niankai. Unsupervised commonsense question answering based on multi-granularities [J]. Journal of North University of China(Natural Science Edition)，2026，47(1)：62-70.

Unsupervised Commonsense Question Answering Based on Multi-Granularities

YANG Zhizhuo，WANG Niankai

（School of Computer and Information Technology，Shanxi University，Taiyuan 030006，China）

Abstract：As a natural language understanding task，commonsense question answering（CQA）is significantly more challenging than conventional question answering tasks. It requires the model to possess stronger commonsense reasoning capabilities. Currently，unsupervised methods for CQA have achieved relatively good performance on several datasets，but these approaches struggle to adequately mine and utilize commonsense knowledge，limiting the model's reasoning ability in complex scenarios. To address this issue，this paper proposed a novel unsupervised CQA method，whose core advantage lay in effectively integrating external commonsense knowledge through unsupervised learning，thereby enhancing the model's generalization capability and reasoning depth. Firstly，the method classifies questions into scientific commonsense questions and everyday event questions. Then，it generates corresponding knowledge

收稿日期：2025-03-24

作者简介：杨陟卓(1983—)，男，副教授，博士，主要从事自然语言处理、信息检索及数据挖掘等方面的研究。E-mail：yangzhizhuo@sxu.edu.cn。

prefixes based on the question type. Next, these knowledge prefixes are input into a pre-trained language model to produce multi-granularities commonsense knowledge through large model prompts. Finally, the multi-grained knowledge is leveraged to assist the answer generation module in reasoning. The adoption of an unsupervised approach not only reduces the reliance on annotated data but also better adapts to diverse commonsense scenarios, demonstrating its flexibility and generalizability in practical applications. Experimental results show that the proposed method significantly outperforms baseline models on relevant datasets, validating its correctness and rationality in unsupervised CQA tasks.

Key words: commonsense question answering; large model prompting; knowledge generation; answer reasoning

0　引　言

常识性问答（CQA）作为一项比传统问答任务更具挑战性的自然语言理解任务，难点在于需要额外的常识性知识来辅助推理，而这些知识往往无法直接从给定的上下文中获取。由于其在虚拟助手、社交聊天机器人等领域的广泛应用前景，以及涉及的知识挖掘与表示、语言理解与计算、答案推理与生成等关键科学问题，常识问答受到了工业界和学术界的广泛关注。现有研究主要通过利用外部常识知识库来提升常识问答的性能[1-3]。Zhao等[4]提出的大语言模型与设计模块的迭代交互实现参数高效微调是间接将常识知识注入语言模型的方法。另一类方法是将问题置于知识图中并基于图结构进行推理。例如：Zhang等[5]将上下文信息融入知识图；Wang等[6]将上下文表示与知识表示分离，仅在最后阶段进行联合答案推理；Sun等[7]则双向融合上下文与知识图信息，生成融合知识信息的查询表示 Q^N 和融合问题信息的图表示 X^N，进而联合推理答案。尽管这些方法能够显著提升性能，但其在构建或寻找合适知识库的过程中往往消耗大量资源。还有一类方法是将预训练语言模型作为常识知识的来源。例如：Liu等[8]提出用提示来生成知识；Sun等[9]提出两阶段生成提示，第一阶段生成知识，第二阶段生成答案；Chen等[10]提出利用生成式语言模型生成对比解释；而Wang等[11]提出利用合成数据模仿大语言模型生成基本原理，Liu等[12]提出利用强化学习优化知识生成。除上述方法之外，Niu等[13]提出基于语义的无监督常识问答方法，Shwartz等[14]提出基于自对话的无监督常识问答方法，Wang等[15]提出的全能思考者等方法则无需外部知识库，完全依赖无监督方式进行常识问答。

尽管现有方法不依赖于大规模领域内或领域外的标注数据，并能以通用方式生成高质量知识，但是这些工作忽视了不同常识表达方式在回答常识问题中的作用，从而导致模型对问题和篇章的理解不够充分。表 1 为问题分类示例，可以看出问题被分为科学常识和日常事件问题。表 1 为一个常识问答示例，针对该问题，可以生成 2 种不同粒度的知识（规则和记叙），其中，规则是句子级别的知识，而记叙是篇章级别的知识。可以看出，面对日常事件问题时，生成的规则仅在前 4 条提供了有用的知识，并且还包含了不正确的选项“classroom”；生成的记叙则详细描述了一个成年人在办公室使用胶棒的情景。如果问答系统能够基于问题生成更加适合的知识类型，就可以帮助系统更加准确地回答问题。

表 1　常识问答任务示例

Tab. 1　An example of CSQA task

Question: Where do adults use glue sticks? **Answer**: **office**
Commonsense Rules: 1.Glue sticks are commonly used for various arts and crafts projects. 2.Adults often engage in arts and crafts activities at home. 3.Glue sticks are commonly found in places where arts and crafts supplies are kept. 4.Adults may also use glue sticks in professional settings, such as an **office** or classroom. **Narrative**: When Jane recalled her past experience with glue sticks, she vividly remembered a time when she was working in an **office**. She was tasked with putting together a presentation board for a client meeting. …… Jane grabbed a glue stick and used it to attach picturesand charts onto the presentation board. The glue stick proved to be quick and mess－free, allowing her to complete the task efficiently.

受提示学习的启发，本文提出了一种利用提示样例和提示模板引导大语言模型生成高相关性的多粒度知识的通用方法。该方法首先对问题进行分类，判断其属于日常事件问题还是科学常识问题；其次，根据不同类型的问题生成相应的知识前缀；随后，将得到的知识前缀输入 GPT-2

中，生成不同粒度的知识；最后，利用得到的不同粒度的知识辅助问答推理模块进行答案生成。该方法沿用了ArT[15]的知识模板，所不同的是本文除了规则模板外，还提出了记叙模板，通过判断问题类别，利用知识模板引导大模型生成与问题类型匹配的知识形式。同时，本文利用提示样例和提示模板诱导大模型进行知识生成，使生成的知识与问题相关性更高。在多个相关数据集上的实验结果表明，本文方法的性能显著优于所有无监督基线方法，验证了其有效性和优越性。

1 本文方法

本研究聚焦于多选形式的无监督常识问答任务，该任务由一个上下文C和相关问题Q组成，要求模型从给定的选项集$O=\{O^i\}_{i=1}^{|O|}$中选择正确答案。所提方法的模型框架如图1所示，主要包括3个模块：1）问题分类模块用于对问题进行分类；2）知识生成模块，用于根据不同类型的问题生成相应的知识前缀，将得到的知识前缀输入GPT-2中，生成完整的记叙或规则；3）推理模块，用于预测问题的答案。

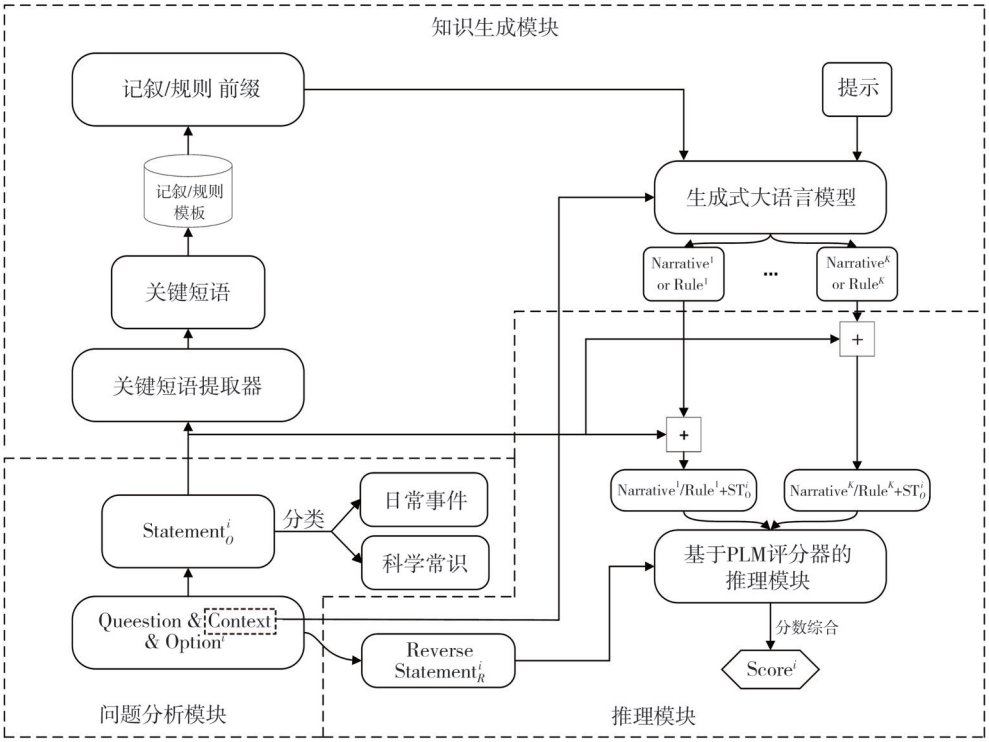

图1 本文所提模型框架示意图

Fig. 1 Framework of the model proposed in this paper

1.1 问题分类模块

该模块以*Option*(O^i)及其对应的*Context*(C)和*Question*(Q)为输入(如图1问题分类模块所示)，首先将原始疑问句Q重写为陈述形式，然后将$\langle C, Q, O^i\rangle$拼接为语句*Statement*$_O^i$(ST_O^i)。通过研究发现，日常事件常识通常较为隐晦，更适合通过记叙这种粗粒度知识来表达；而科学常识则通常以规则形式组织，问题和选项中常包含科学术语，更适合通过规则这种细粒度知识来表达。本文设计了6个提示样例和1个提示模板(如表2所示)，将正确答案对应的ST_O^i填入提示模板，形成完整的提示模板，判断ST_O^i为日常事件(Daily Event)还是科学常识(Scientific Commonsense)。

如表3所示，疑问句“What is the effect of this?”和“What is the cause of this?”对应的陈述形式分别为“As a result，”和“Because”。拼接而成的语句ST_O^i分别为“The grape juice fermented. As a result，the juice turned to wine”和“The man got a discount on his groceries. Because he used a coupon”。这两个句子分别为科学常识和日常事件。

1.2 知识生成模块

为了充分提高GPT-2利用常识的效果，克服ArT的不足，本文利用前缀生成模块，以简单通用的方式生成不同粒度的自然语言常识描述(在

本文的工作中定义为“记叙”或“规则”）。前缀生成分为2个步骤：关键短语提取和记叙或规则生成。如图1中知识生成模块所示。

表2 提示样例和提示模板

Tab. 2 Prompt examples and prompt template

Determine whether this is a matter of scientific commmonsense or a daily event based on the ST_O^i provided. Examples:
ST_O^i: My body cast a shadow over the grass because the sun was rising. Class: daily event ST_O^i: Marine life diminished because oil spilled into the ocean. Class: scientific commonsense ST_O^i: Our group's conversation gradually lulled to silence.As a result, I felt awkward. Class: daily event ST_O^i: The magnet attracted the paperclip.As a result, the paperclip stuck to the magnet. Class: scientific commonsense ST_O^i: I excused myself from the group because my phone rang. Class: daily event ST_O^i: Air pollution in the city worsened because factories increased their production. Class: scientific commonsense ST_O^i:{$Statement_O^i$} Class:

表3 问题分类及结果示例

Tab. 3 Question classification and system outputs examples

上下文与选项	问题	问题类型	ArT	本文模型
Context: The grape juice fermented. Options: A.The juice turned to wine. B.The juice evaporated.	Q1:What is the effect of this?	科学常识	B （错误）	A （正确）
Context：The man got a discount on his groceries. Options: A.He greeted the cashier. B.He used a coupon.	Q2:What is the cause of this?	日常事件	A （错误）	B （正确）

1.2.1 关键词提取

对于1.1节提到的语句，该阶段使用无监督的关键短语提取器从ST_O^i中提取关键短语。具体来说，该阶段提取了3种类型的短语：名词短语（Noun Phrase，NP）、动词短语（Verb Phrase，VP）和人名短语（Person Name Phrase，PNP）。

为实现此目标，本文为每种短语设计了简单的上下文无关语法（Context-Free Grammar，CFG）规则

$$\cdot NP \rightarrow (nn|adj)^{*} + nn,$$

$$\cdot VP \rightarrow vb + (pr)\{0,1\} + NP,$$

$$\cdot PNP \rightarrow (pn)\{1,2\},$$

其中，VP、NP、PNP为非终止子；vb（动词）、nn（名词）、adj（形容词）、pn（人名）和pr（介词）是限定词；‘+’表示串联；$\{a, b\}$表示重复次数范围从a到b；‘*’相当于$\{0, \infty\}$。将这些CFG规则添加到自然语言工具包（Natural Language ToolKit）[16-17]的RegexpPaser工具中，提取前5个最重要的关键短语，无需标签或微调模型，本文的词嵌入来自预训练语言模型Embeddings from Language Models[18]。

1.2.2 记叙或规则生成

在获得关键短语后，需要根据问题类型检索记叙或规则模板集，生成记叙或规则前缀。若问题为日常事件，生成记叙前缀，考虑以下情况：1）对于一个对象（NP），思考“过去与对象（NP）的一段经历”；2）对于一种行为（VP），思考“该行为（VP）发生后的具体经历”，“该行为（VP）发生前的具体经历”或“与该行为（VP）相关的具体经历”；3）对于一个人（PNP），思考“对他的印象”或“他做某事的动机/感受/反应相关的具体经历”。本文提出的记叙模板集如表4所示。若问题为科学常识，生成规则前缀，考虑以下情况：1）对于一个对象（NP），思考“它是什么”和“它的用途是什么”；2）对于一种行为（VP），思考“它意味着什么”和“它发生前后怎样”；3）对于一个人（PNP），思考“他/她是谁”以及“他/她的感受/动机/反应是什么”。本文提出的规则模板集如表5所示。使用关键字类型作为键，记叙或规则模板列表作为值，构建查找表。通过查找表，可以快速检索关键短语对应的模板，并用关键短语替换标签（[NP]、[VP]和[PNP]），形成记叙或规则前缀。虽然查找表结构简单，但是该表在不同基准测试中表现出良好的有效性和通用性。

表4 记叙模板查找表

Tab. 4 Narrative templates lookup table

Key	Value	Replacing rule
“NP”	{“When Jane recalled her past experience with [NP],”}	directly replace
“VP”	{“Jane remember a specific experience after [VP],”,“Jane remember a specific experience before [VP],”,“Jane remember a specific experience about [VP], ”}	convert to gerund first
“PNP”	{“When Jane recalled her impression on [PNP],[PNP] is a,”,“When Jane recalled a specific experience with [PNP],[PNP] did this because,”,“When Jane recalled a story about [PNP],[PNP] felt”}	directly replace

表5 规则模板查找表

Tab. 5 Rule templates lookup table

Key	Value	Replacing rule
“NP”	{“The definition of [NP] is”, “The main function of [NP] is”, “[NP] is a/an”}	directly replace
“VP”	{“[VP] means”, “After [VP], ”, “Before [VP], ”}	convert to gerund first
“PNP”	{“[PNP] is a/an”, “[PNP] felt”, “After this, [PNP]”, “[PNP] did this because”}	directly replace

对于每个记叙或规则前缀，该模块将其与上下文连接，并输入到GPT-2中，通过提示生成完整的记叙或规则。具体采用p=0.8核采样[19]作为解码策略，以增加生成文本的多样性。同时，为保证生成内容的质量和数量，该模块根据GPT-2估计的困惑度对所有生成的记叙或规则进行排序，保留前K个记叙或规则，构建记叙或规则集：$\{narrative^k\}_{k=1}^{K}/\{rule^k\}_{k=1}^{K}$。

1.3 推理模块

每个记叙或规则将插入ST_O^i作为额外的知识，辅助后续选项评分，如图1推理模块所示。句子似然度是选项O^i评分常用的函数。

$$score_O^{i,k}=P_{LM}(O^i|narrative^k+ST_O^i-O^i)=\frac{1}{|O^i|}\sum_{t=1}^{|O^i|}\log P_{LM}(O_t^i|narrative^k+ST_O^i-O^i+O_{<t}^i),\tag{1}$$

$$score_O^{i,k}=P_{LM}(O^i|rule^k+ST_O^i-O^i)=\frac{1}{|O^i|}\sum_{t=1}^{|O^i|}\log P_{LM}(O_t^i|rule^k+ST_O^i-O^i+O_{<t}^i),\tag{2}$$

$$Score_O^i=\frac{1}{K}\sum_{k=1}^{K}Score_O^{i,k}。\tag{3}$$

式(1)、式(2)表示结合$narrative^k$/$rule^k$和ST_O^i，计算去除自身O^i后，仍能生成O^i的概率。式(3)表示记叙或规则对选项O^i的综合分数。

对于因果推理题，该模块引入逆向思维(RT)，进行因果之间的双向推断。除了1.1节中的有序重写(ST_O^i)，本文还应用反向重写，将$\langle O^i,Q^R,C\rangle$按顺序连接(记为ST_R^i)，其中Q^R为Q的反义形式。具体而言，在因果推理任务中，“Because”和“Therefore”是两个相反的问题。为了进行双向推理，除了有序推理$C+Q\rightarrow O^i$的$Score_O^i$外，本文设置了反向推理评分函数$Score_R^i$：$O^i+Q^R\rightarrow C$为

$$Score_R^i=P_{LM}(C|ST_R^i-C)=\frac{1}{|C|}\sum_{t=1}^{|C|}\log P_{LM}(C_t|ST_R^i-C+C_{<t})。\tag{4}$$

为了利用双向推断的优势，本文设计了一个混合评分函数

$$Score_X^i=\frac{1}{2}(Score_O^i+Score_R^i)。\tag{5}$$

默认情况下最终预测答案($\hat{A}$)选择为

$$\hat{A}=\underset{O^i}{\operatorname{argmax}}\ Score_O^i。\tag{6}$$

对于因果推理题，只需把式(6)中的$Score_O^i$换成$Score_X^i$。选取双向推断分数更高的选项作为最终答案。

2 实 验

2.1 数据集

为了验证本文所提模型的有效性，选取3个不同的常识性QA数据集COPA(Choice of Plausible Alternative)、SocialIQA(Social Interaction Question Answering)和SCT(Story Cloze Test)[20-22]作为实验数据。

1) COPA：评估对某一事件的因果推理能力。每个问题都有两个候选选项。

2) SocialIQA：评估社交互动的推理能力。问题类型多样，包括主体的动机、反应、个性等。每个问题提供3个候选项。

3) SCT：要求模型从两个选项中选择给定短篇故事的正确结局。每个故事由4个句子组成。

由于3个数据集的两个测试集未公开，本文报告了开发集的所有结果。需要注意的是，标签信息保持不可见，仅用于最终精度评估。

2.2　基线模型和对比模型

实验的基线仅使用预训练语言模型作为评分器，未注入任何显式知识。采用第1.3节所述的公式（3）作为评分函数。同时，本文将所提方法与其他先进的无监督模型进行了对比，具体模型如下：

1）Self-talk[14]：通过GPT-2的两阶段提示获取知识。不同任务需设计特定问题前缀。

2）SEQA[13]：应用GPT-2生成数百个伪答案，并将其与每个选项进行比较。然而，它的评分者依赖于在大规模标记的自然语言推理数据集上微调的SRoBERTa$_{\text{large}}$，这并不是严格无监督的方法。为了公平比较，本文设计了另一种设置，用原始设置（仅在未标记的文本上进行预训练）替换微调的SRoBERTa$_{\text{large}}$。

3）CGA：采用一个生成型知识库COMET，该知识库是在现有种子知识库（例如ConceptNet），构建上下文相关知识图来进行推理。

4）ArT[15]：利用精心设计的通用注释模板，并通过GPT-3.5 turbo来生成全方位的知识，笔记前缀都是通用的，不需要为不同的任务专门设计。这是严格无监督的方法。

2.3　设　置

本文使用单块Tesla V100进行实验，Pytorch版本为2.5.0。根据之前的工作，实验采用OpenAI GPT作为预训练语言模型的骨干。为了获得可靠和可重复的结果，本文在4种不同规模的GPT-2[23]上进行了知识评估。对于Self-talk和ArT，在知识生成和选项评分过程中应用了相同规模的GPT-2。对于SEQA，不同规模的GPT-2用于伪答案生成，SRoBERTa$_{\text{large}}$用于语义相似度计算。为了区分，SRoBERTa$^{\text{NLI}}_{\text{large}}$和SRoBERTa$^{\text{Origin}}_{\text{large}}$分别指对NLI数据集进行进一步微调和不进行进一步微调的SRoBERTa$_{\text{large}}$。对于Self-talk和SEQA，本文使用其原始设置重新运行代码，并报告本文重新运行的结果和来自其出版物的结果。对于CGA，本文报告了Niu等（2021）提供的结果。对于ArT，本文使用其原始设置重新运行代码，并报告重新运行的结果。默认情况下，注释集（K）的大小设置为32。除了$Score^i_O$，ArT在COPA上采用了另一个设置$Score^i_X$。对于本文所提出的方法，除了规则前缀，本文增加了记叙前缀。GPT-3.5 turbo用于知识生成，不同规模的GPT-2用于选项评分。记叙集（N）的大小根据不同数据集设置，对于COPA数据集，记叙集（N）的大小设置为7；对于SocialIQA数据集，记叙集（N）的大小设置为14；对于SCT数据集，记叙集（N）的大小设置为14。规则集（R）的大小都设置为32。除了$Score^i_O$，本文所提出的方法在COPA上也采用了另一个设置$Score^i_X$。

2.4　实验结果

2.4.1　对比实验

实验结果如表6所示，可以看出，在所有数据集上，本文提出的方法与其他完全无监督的模型相比，获得了最优的性能。

表6　模型实验结果

Tab.6　Model experimental results　%

Dataset	Models	Our (re-)running				Published
		DistillGPT-2	GPT-2$_{\text{medium}}$	GPT-2$_{\text{large}}$	GPT-2$_{\text{xlarge}}$	GPT-2$_{\text{xlarge}}$
COPA	Baseline	57.8	62.4	65.8	66.0	—
	SEQA	51.4(63.0)	53.0(68.4)	53.8(72.0)	54.4(75.4)	79.4
	Self-talk	59.8(↑2.0)	65.0(↑2.6)	66.6(↑0.8)	66.2(↑0.2)	68.6
	CGA	—	—	—	—	72.2
	ArT	60.2(↑2.4)	64.8(↑2.4)	67.0(↑1.2)	67.6(↑1.6)	—
	ArT(Scorei X)	61.0(↑3.2)	65.6(↑3.2)	69.4(↑3.6)	69.8(↑3.8)	—
	Our	**62.4**(↑4.6)	**67.2**(↑4.8)	**70.6**(↑4.8)	**70.0**(↑4.0)	—
SocialIQA	Baseline	41.3	44.3	45.5	45.9	—
	SEQA	34.9(43.9)	35.9(44.6)	36.5(46.6)	36.6(47.5	47.5
	Self-talk	40.5(↓0.8)	44.8(↑0.5)	46.1(↑0.6)	47.2(↑1.3)	47.5
	CGA	—	—	—	—	45.4
	ArT	42.0(↑0.7)	45.6(↑1.3)	47.6(↑2.1)	47.3(↑1.4)	—
	Our	**43.4**(↑2.1)	**47.9**(↑3.6)	**49.2**(↑3.7)	**49.0**(↑3.1)	—
SCT	Baseline	59.6	67.4	69.1	70.5	—
	SEQA	50.7(74.7)	53.3(80.5)	54.2(82.4)	54.9(83.2)	83.2
	Self-talk	59.8(↑0.2)	68.5(↑1.1)	69.2(↑0.1)	70.4(↓0.1)	70.4
	CGA	—	—	—	—	71.5
	ArT	60.2(↑0.6)	68.3(↑0.9)	69.5(↑0.4)	71.6(↑1.1)	—
	Our	**60.6**(↑1.0)	**68.5**(↑1.1)	**69.7**(↑0.6)	**72.2**(↑1.7)	—

此外，值得注意的是，在COPA数据集上，本文方法使用GPT2-xlarge的评估结果比GPT2-large还低0.6%。这表明，仅使用增加模型参数的方法来提高语言模型作为句子评分器的性能可能存在局限性。与本文的方法相比，ArT无法利用不同粒度的常识知识来回答不同类型的问题，其准确性略低于本文提出的方法。这表明对不同类型的问题采用不同粒度的知识可以提高回答的准确性。与ArT相比，Self-talk无法保持有效性，其准确性有时甚至略低于基线。这表明Self-talk产生的知识可能存在噪声，从而会误导模型推理。在SRoBERTa$^{\mathrm{NLI}}_{\mathrm{large}}$的帮助下，SEQA可以在所有数据集上取得满意的结果，尤其是在SCT上(超过所有模型的10%以上)。

2.4.2 消融实验

本文深入分析了记叙集的大小i_N以及最大tokens的限制对系统性能的影响。

本文在COPA上对记叙集的大小进行研究，并分析了记叙集的大小对实验结果的影响，结果如表7所示，其中i_N表示记叙集的大小为i，可以看出，随着记叙集大小i_N的增加，系统性能得到提升，表明系统可以通过增强挖掘大语言模型所产生的知识，对回答问题形成更好的理解。但是，当记叙集i_N超过一定值之后，系统性能反而降低，表明更多的记叙会带来一定的噪声。本文研究并分析了最大tokens的限制对实验结果的影响。实验结果如图2所示，其中，7_N表示记叙集的大小，都设置为7，120Max_tokens和150Max_tokens分别表示生成的记叙最大tokens大小不能超过120和150。可以看出，当最大tokens限制为150时，4种不同规模的GPT-2评估下系统性能更好。

表7 记叙集大小对系统性能的影响

Tab. 7 The influence of narrative size on the system's performance

%

i^N	GPT-2_{distil}	GPT-2_{medium}	GPT-2_{large}	GPT-2_{xlarge}
1_N	59.8	66.6	67.2	68.4
2_N	61.6	66.6	67.6	68.4
3_N	61.8	66.8	68.2	68.8
4_N	61.8	67.0	69.0	68.8
5_N	62.6	67.2	69.8	68.8
6_N	62.8	68.0	70.0	70.2
7_N	**63.4**	**68.0**	**71.8**	**70.4**
8_N	61.0	67.6	68.2	70.2

为了确定系统性能增长的来源，本文在均使用GPT-3.5 turbo作为知识生成模型的情况下，将本文的方法与ArT进行比较，具体实验结果如图3所示，值得注意的是本文的方法在GPT-2_{distil}和GPT-2_{large}上的性能要远高于ArT，而在GPT-2_{medium}和GPT-2_{xlarge}上的性能略低于ArT。总体而言，本文所提出的方法要优于ArT。

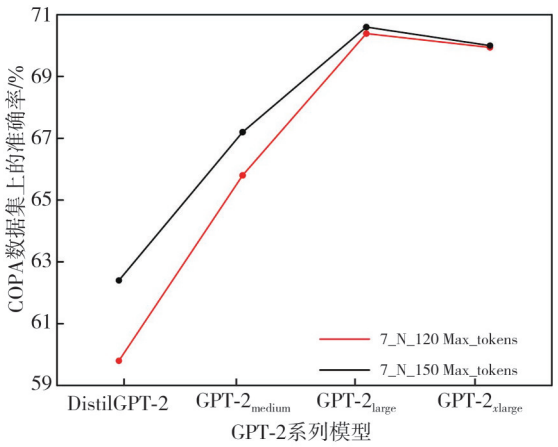

图2 最大tokens的限制对系统性能的影响

Fig. 2 Influence of the maximum token limit on the system's performance

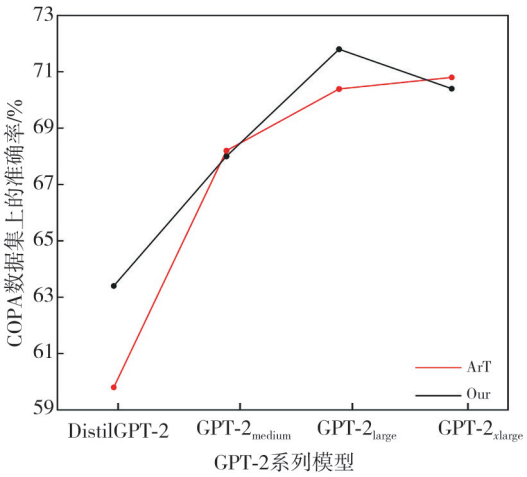

图3 知识生成模型对系统性能的影响

Fig. 3 Influence of the knowledge generation model on the system's performance

2.4.3 示例分析

表3给出部分示例，对于这些示例本文模型可以正确回答，而ArT模型回答错误。可以看出，通过利用不同粒度的常识知识来回答不同类型的问题，系统灵活调整了常识表达方式，有助于更好地理解常识问题，因此本文的模型给出了正确答案。

3 结 论

本文提出了一种基于多粒度知识的无监督常识问答方法，该方法首先分析上下文中的关键信息，随后在此基础上以联想方式生成高度相关且粒度多

样的知识。随后引入问题分类模块，以便根据不同类型的问题采用相应粒度的知识进行回答。最后，在COPA、SocialIQA和SCT三个基准数据集上对所提出的方法进行了测试。实验结果表明，本文方法在性能上超越了先前的无监督模型，并在以GPT-2为主干的所有规模上均展现出稳定的性能。

为进一步提升模型在复杂常识推理场景中的适应性，未来工作计划从两方面改进：1）知识表示扩展方面，引入因果图、场景脚本等结构化知识表示，构建多粒度知识融合框架，增强对交互性推理任务的支持。2）架构优化方面，设计知识-推理并联交互架构，通过共享注意力机制实现知识生成与推理评分的双向协同，其中动态路由模块将根据问题类型自动选择最优知识组合。该方案预计采用多任务学习联合优化知识生成与推理目标，以提升端到端的推理效率。实验将重点验证扩展知识类型对开放域推理任务的增益效果。

利益冲突声明/Conflict of Interests

所有作者声明不存在利益冲突。

All authors declare no relevant conflict of interests.

参考文献：

[1] WANG W Q, FANG T Q, DING W X, et al. CAR: conceptualization-augmented reasoner for zero-shot commonsense question answering [C]//Conference on Empirical Methods in Natural Language Processing, 2023: 13520-13545.

[2] HE J, GUTIÉRREZ-BASULTO V, PAN J Z. BUCA: A binary classification approach to unsupervised commonsense question answering [C]//61st Annual Meeting of the Association for Computational Linguistics, 2023: 376-387.

[3] SHI H, WANG W, FANG T, et al. QADYNAMICS: Training dynamics-driven synthetic QA diagnostic for zero-shot commonsense question answering [C]//Findings of the Association for Computational Linguistics: EMNLP 2023, 2023: 15329-15341.

[4] ZHAO Z, HU L, ZHAO H, et al. Knowledgeable parameter efficient tuning network for commonsense question answering [C]//Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics, 2023: 9051-9063.

[5] ZHANG L, LI R. KE-GCL: Knowledge enhanced graph contrastive learning for commonsense question answering[C]//Findings of the Association for Computational Linguistics: EMNLP 2022, 2022: 76-87.

[6] WANG Y, ZHANG H, LIANG J, et al. Dynamic heterogeneous-graph reasoning with language models and knowledge representation learning for commonsense question answering [C]//61st Annual Meeting of the Association for Computational Linguistics, 2023: 14048-14063.

[7] SUN Y, SHI Q, QI L, et al. JointLK: Joint reasoning with language models and knowledge graphs for commonsense question answering [C]//2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2022: 5049-5060.

[8] LIU J, LIU A, LU X, et al. Generated knowledge prompting for commonsense reasoning [DB/OL]. (2021-10-15) [2024-07-20]. https://arxiv. org/abs/2110. 08387.

[9] SUN Y, ZHANG Y, QI L, et al. TSGP: Two-stage generative prompting for unsupervised commonsense question answering [C]//Findings of the Association for Computational Linguistics: EMNLP 2022: 968-980.

[10] CHEN Q, XU G, YAN M, et al. Distinguish before answer: Generating contrastive explanation as knowledge for commonsense question answering [C]//Findings of the Association for Computational Linguistics: ACL 2023, 2023: 13207-13224.

[11] WANG C, CAO P, LI J, et al. Leros: Learning explicit reasoning on synthesized data for commonsense question answering [C]//2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), 2024: 10303-10315.

[12] LIU J, HALLINAN S, LU X, et al. Rainier: Reinforced knowledge introspector for commonsense question answering [C]//Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, 2022: 8938-8958.

[13] NIU Y L, HUANG F, LIANG J M, et al. A semantic-based method for unsupervised commonsense question answering [C]//59th Annual Meeting of the Association for Computational Linguistics, 2021: 3037-3049.

[14] SHWARTZ V, WEST P, LE BRAS R, et al. Unsupervised commonsense question answering with self-talk [C]//2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2020: 4615-4629.

[15] WANG J, ZHAO H. ArT: All-round thinker for unsupervised commonsense question answering[C]// 29th International Conference on Computational Linguistics,2022:1490-1501.
[16] CHOMSKY N, SCHÜTZENBERGER M P. The algebraic theory of context-free languages [M]. Amsterdam: Elsevier, 1959.
[17] BIRD S. NLTK: The natural language toolkit[C]// Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions, 2006:69-72.
[18] PETERS M E, NEUMANN M, IYYER M, et al. Deep contextualized word representations[C]//2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2018: 2227-2237.
[19] HOLTZMAN A, BUYS J, DU L, et al. The curious case of neural text degeneration [C]//International Conference on Learning Representations 2020, 2020: 1-16.
[20] ROEMMELE M, BEJAN C A, GORDON A S. Choice of plausible alternatives: An evaluation of commonsense causal reasoning[C]//AAAI Spring Symposium: Logical Formalizations of Commonsense Reasoning, 2011: 90-95.
[21] SAP M, RASHKIN H, CHEN D, et al. SocialIQa: Commonsense reasoning about social interactions[C]// 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), 2019: 4463-4473.
[22] MOSTAFAZADEH N, CHAMBERS N, HE X, et al. A corpus and cloze evaluation for deeper understanding of commonsense stories[C]//2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language, 2016: 839-849.
[23] RADFORD A, WU J, CHILD R, et al. Language models are unsupervised multitask learners[EB/OL]. (2019-02- 14)[2024-07-17]. https://www.semanticscholar.org.

(上接第52页)

[13] BYUN J H. The analytical characteriza-tionof 2-D braided textile composites[J]. Composites Science and Technology, 2000, 60(5): 705-716.
[14] TABIEI A, YI W T. Comparative study of predictive methods for woven fabric composite elastic properties [J]. Composite Structures, 2002, 58(1): 149-164.
[15] 吴宁, 解锡明, 杨洁. 经密对2.5D碳纤维织造损伤影响的实验评价[J]. 天津工业大学学报, 2020, 39(3): 15-21. WU Ning, XIE Ximing, YANG Jie. Experimental evaluation for effect of warp density on 2.5D carbon fiber weaving damage[J]. Journal of Tianjin Polytechnic University, 2020, 39(3): 15-21.(in Chinese)