基于全局覆盖机制与表示学习的生成式知识问答技术

刘琼昕[1, 2] 王亚男[2] 龙 航[2] 王佳升[2] 卢士帅[2]

摘 要 针对现有生成式问答模型中陌生词汇导致答案准确率低下的问题和模式混乱导致的词汇重复问题, 本文提出引入知识表示学习结果的方法提高模型识别陌生词汇的能力, 提高模型准确率. 同时本文提出使用全局覆盖机制以平衡不同模式答案生成的概率, 减少由预测模式混乱导致的重复输出问题, 提高答案的质量. 本文在知识问答模型基础上结合知识表示学习的推理结果, 使模型具备模糊回答的能力. 在合成数据集和现实世界数据集上的实验证明了本模型能够有效地提高生成答案的质量, 能对推理知识进行模糊回答.

关键词 生成式知识问答, 覆盖机制, 知识表示学习, 自然语言处理, 深度学习

引用格式 刘琼昕, 王亚男, 龙航, 王佳升, 卢士帅. 基于全局覆盖机制与表示学习的生成式知识问答技术. 自动化学报, 2022, **48**(10): 2392−2405

DOI 10.16383/j.aas.c190785

Generative Knowledge Question Answering Technology Based on Global Coverage Mechanism and Representation Learning

LIU Qiong-Xin[1, 2] WANG Ya-Nan[2] LONG Hang[2] WANG Jia-Sheng[2] LU Shi-Shuai[2]

Abstract Aiming at the problem of low answer accuracy caused by unfamiliar words in the existing generative question answering model and the problem of vocabulary repetition caused by pattern confusion, this paper proposes a method of introducing knowledge representation learning results to improve the model′s ability to recognize unfamiliar words and improve the accuracy of the model. At the same time, this paper proposes to use a global coverage mechanism to balance the probability of answer generation in different modes, reduce the repeated output problem caused by the confusion of prediction modes, and improve the quality of the answer. Based on the knowledge question answering model, this paper combines the inference results of knowledge representation learning, so that the model has the ability to answer fuzzy answers. Experiments on synthetic datasets and real-world datasets demonstrate that this model can effectively improve the quality of generated answers and can provide fuzzy answers to reasoning knowledge.

Key words Generative knowledge base question answering, coverage mechanism, knowledge representation learning, natural language processing (NLP), deep learning

Citation Liu Qiong-Xin, Wang Ya-Nan, Long Hang, Wang Jia-Sheng, Lu Shi-Shuai. Generative knowledge question answering technology based on global coverage mechanism and representation learning. *Acta Automatica Sinica*, 2022, **48**(10): 2392−2405

对于用户用自然语言提出的问题, 知识问答系统 (Knowledge base question answering, KBQA) 通常提供短语和实体形式的精确答案[1]. 在现实环境中, 人们希望答案能够使用自然语言回答提出的问题, 这需要答案用完整/部分自然语言句子而不是单个实体/短语表达, 系统不仅需要解析问题, 还要从 KB (Knowledge base)[2] 检索相关事实, 生成一个适当的回复.

生成式知识问答任务使用 Seq2Seq[3] 框架来实现使用自然语言回答提出的问题, 不同于其他问答模型, 生成式知识问答模型无需其他自然语言处理工具, 可以在端到端的框架中同时实现分析问题, 从知识库检索事实, 并且生成正确、连贯以及自然的答案.

实现生成一句连贯的自然语言答案会面临很多的挑战, 比如词典外词汇 (Out of vocabulary, OOV) 问题: 由于模型词典大小有限, 在构建词典时, 会忽略掉一些词频较低的词汇, 当编码器端输入 OOV

收稿日期 2019-11-15 录用日期 2020-04-10
Manuscript received November 15, 2019; accepted April 10, 2020
国家自然科学基金 (62072039) 资助
Supported by National Natural Science Foundation of China (62072039)
本文责任编委 张民
Recommended by Associate Editor ZHANG Min
1. 北京市海量语言信息处理与云计算应用工程技术研究中心 北京 100081 2. 北京理工大学计算机学院 北京 100081
1. Beijing Engineering Research Center of High Volume Language Information Processing and Cloud Computing Applications, Beijing 100081 2. School of Computer Science and Technology, Beijing Institute of Technology, Beijing 100081

词时会用 “UNK (Unknown)”来代替, 但这样做很有可能导致输出端也输出 “UNK”, 损失了原有的词义信息.

文献 [4] 提出 CopyNet 复制网络, 文献 [5] 提出指针网络来缓解 OOV 问题, 二者思路类似, 即构建一个由源端 OOV 词构成的词汇表, 当预测目标端的词时, 会输出源端词表和现有词表中的词汇, 减小 “UNK”词出现的概率. 文献 [6] 结合了之前的生成式问答模型和 CopyNet 网络, 提出的 CoreQA 模型在知识库问答任务上能够以自然的方式回答复杂问题.

尽管上述工作在生成式问答任务上有了很大进展, 但是仍存在以下不足:

1) 生成式问答模型的模式混乱问题. 模型可能会生成答案如 Q: “鲁迅的原名是什么”, A: “原名周树人周树人周树人”的情况, 词汇重复降低了答案质量. 导致此类问题的原因是从问题端或知识库复制词汇到答案的过程中, 注意力机制引发了各个模式的混乱, 模型往往会陷入某种模式中无法跳出, 导致答案的可读性下降.

2) 陌生词汇问题. 尽管引入 CopyNet 能够缓解从问题引入 OOV 词汇的问题, 但是知识问答任务还需要引入外部的知识库, 模型通过问题识别出关系, 拷贝知识库对应事实的尾实体填补到答案中. 陌生关系被识别为 OOV 词, 导致无法寻找到正确知识.

另外查询到的知识还要指导基础词典中词汇的生成. 举例如 Q: “小明在哪个城市生活”, A: “他在北京生活”. 知识库中存在事实三元组 (小明, 性别, 男), 如果基础词典中没有描述性别的词汇, 对于模型来说 “男”和 “女”会被识别为 ⟨unk⟩, 无法判断是使用 “他”还是 “她”作为实体的代词. 构成上述问题的主要原因是陌生词汇用 ⟨unk⟩ 代替后向量表征不唯一.

3) 当用户的问题在图谱中没有相关的知识作为支撑时, QA (Question answering) 系统通常会答非所问, 或者回答错误的答案.

针对模式混乱问题, 本文通过全局覆盖机制来控制 3 个模式的切换, 当某个模式关注度足够高的时候, 提高其他模式的受关注的概率, 控制模式间的切换.

针对生成式问答模型中的陌生词汇问题, 本文利用知识表示学习方法生成的实体、关系向量代替基础词典中相应词汇的词向量, 让所有陌生词汇有唯一的向量表征, 提升模型匹配知识的能力.

本文运用知识推理补全知识库中缺失的知识, 模型可以提供给用户推理出的答案, 在构建的生成式知识问答系统基础上通过数据共享的方式, 对推理得到的知识进行模糊问答.

知识表示学习[7−8] 是在知识图谱的构建和利用过程中, 进一步挖掘知识图谱结构信息的方法, 处理的方式是将知识图谱中的实体信息和关系信息映射到低维向量空间, 每一个实体和关系都有其独一无二的向量表征. 类似于自然语言处理 (Natural language processing, NLP) 中的 word2vec[9] 技术, 准确的知识向量表征能够提升相关任务的效果.

综上, 本文提出一种基于表示学习与全局覆盖机制的生成式问答模型 (Multi coverage mechanism over question answering model, MCQA), 能够利用知识图谱信息并使用自然语言回答用户提出的问题. 它具有以下贡献:

1) 针对词汇重复降低答案质量的问题, 提出使用全局覆盖机制来减轻生成模型模式混乱;

2) 提出通过引入知识表示学习结果的方法缓解 OOV 问题, 提高模型答案的准确率;

3) 提出一种知识推理和知识问答的结合方式, 使模型具备模糊回答的能力;

4) 本文在 SimpleQuestion 单关系知识问答数据集、生日问答限定领域数据集和社区问答开放领域数据集上进行实验, 实验结果表明, 与现有方法相比, 本文所提出的模型可以更有效准确地为知识查询问题生成正确、连贯和自然的答案. 结合知识推理方法, 模型已经初步具有分辨原始知识和推理知识的能力, 并能对推理知识进行模糊回答.

1 相关工作

基于知识图谱的问答在自然语言领域拥有很长的历史. 早期的研究主要有 3 种方式: 基于语义分析建模的 KBQA[10]、基于信息抽取建模的 KBQA[11] 以及基于向量建模的 KBQA[12]. 这些传统的 KBQA 方式有很多缺陷, 如需要很多的先验知识、模型复杂、模型不够灵活等, 随着深度学习技术的发展, 基于深度学习的 KBQA 方法已成为研究的重点.

考虑到传统语义解析与 KB 结合不够紧密, Yih 等[13] 提出了查询图的概念. 该模型结合了传统语义分析和深度学习的方法, 语义分析为主导, 在获取相应答案 Pattern 的步骤中, 使用卷积神经网络 (Convolutional neural network, CNN) 选定推导链, 获得了比较显著的成果. Sun 等[14] 提出的 SPARQA (Skeleton-based semantic parsing for question answering) 方法先解析出复杂问句的宏观结构 (Skeleton), 以句子级和单词级进行匹配, 再做后续处理, 取得了不错的实验效果.

考虑到以往只对 Decoder 端改进而忽略了句

法特征, Xu 等[15] 把 Graph2Seq 应用到 KBQA 任务上. Graph2Seq 是一种用于图到序列学习的新的基于注意力的神经网络结构, 对输入图进行编码, 从而通过编码更多的句法信息, 可以提高模型的鲁棒性.

随着深度学习的发展, 出现了一种适用于 KBQA 任务的新型网络模型记忆网络 (Memory network, MMN). 以键值记忆网络[16] 为例, MMN 是一个长期记忆网络, 可以储存大量的先验知识, 可以利用提前构建的知识库提高对话的质量. 对于以知识图谱为知识库的 QA 模型, 键是三元组的头实体和关系, 键值是尾实体. 模型通过倒排索引的方式从知识库中检索出主题词相关的事实, 通过问题的分布式表征和转移矩阵得到查询命令的向量表征, 计算问题向量和每一个键值的相似度得分, 并将相似度 softmax 转化为每一个键值的权重, 读取键值对应的数据值, 也就是问题的答案.

Bordes 等[17] 提出一种基于词嵌入的问答系统, 在记忆网络框架下实现, 并贡献了一个 KBQA 的数据集 SimpleQuestions, 相比较原始的 WebQuestion 数据集, 前者具有更大规模的样本数据, 而且每一个问题的答案只依赖于知识库中的一个三元组. 构建 SimpleQuestions 数据集的初衷在于当前的 KBQA 任务应该优先解决简单问答任务, 而不是直接去解决复杂依赖的问答任务.

上述 KBQA 模型答案反馈给用户的都是简单实体词汇, 为了赋予模型 “说话”的能力, 生成式知识问答任务中使用深度学习中的 Seq2Seq 框架. Seq2Seq 模型广泛应用于诸如机器翻译、文本摘要、对话机器人等领域. 在应用 Seq2Seq 模型的生成式对话中, 词典外词汇 (OOV) 问题不可避免. 为了缓解 OOV 问题而提出了复制网络 (Copy net)[4] 和指针网络 (Pointer net)[5], 他们的基本想法都是当预测目标端词汇时输出源端词表和现有词表中的词汇, 减小 “UNK” 词出现的概率.

Yin 等[18] 提出的 GenQA 模型首先在生成式 KBQA 任务上进行了尝试, 将生成式问答模型与知识问答模型结合, 在目标端生成答案时, 将知识库的知识加入进去. 但是由于在计算问题与事实向量的相似度的时候使用了问题最终的编码, 忽略了问题的时序信息, 导致模型不能处理需要结合多个事实进行回答的情况. He 等[6] 提出了一个 CoreQA 模型, 进一步结合 GenQA 和复制网络, 将每一个问答对所依赖的知识增加到多个, 进一步改善模型生成答案的流畅性和准确性, 并贡献了 CoreQA 的数据集.

Liu 等[19] 从数据入手, 对 CoreQA 模型进行改进, 利用课程学习的思路让模型先从标准易学的数据开始训练, 逐步加大样本难度, 让模型更好地拟合数据.

2 生成式知识问答模型

本模型使用了 Seq2Seq 框架, 模型总体结构如图 1 所示. 模型通过编码器解析问题, 并通过查询知识库中的信息, 使用解码器生成答案. 本节将对模型各个部分的原理和策略进行详细描述.

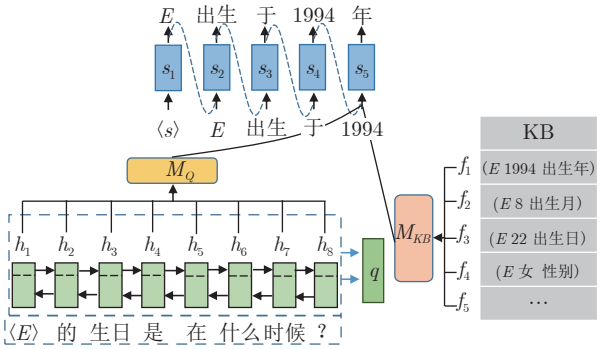


图 1 MCQA 模型图

Fig. 1 The overall diagram of MCQA

智能问答是一个非常复杂的自然语言处理任务, 一个完整的智能问答系统, 首先要求要有完备的知识库作为支撑, 其次要求模型能从知识库中找到正确答案, 并使用完整流畅的一句话反馈给用户[20−21], 当任务背景设置为聊天场景时, 还需要模型具有多轮对话[22−23] 的能力. 所以本文限定部分情景设计相对完备的知识问答模型, 限定场景如下:

1) 单轮生成式知识问答任务, 针对三元组尾实体进行问询.

2) 语料已经通过命名实体识别、主题词识别处理.

3) 给定知识库, 或与主题词相关的知识子图, 以及知识库的表示学习结果.

2.1 词典构成

MCQA 模型根据功能划分为 3 个部分: 生成模式、复制模式、知识库查询模式. 如图 2 所示, 3 种模式有不同的词典, 生成模式的词典是模型的基础词典 V, V 包含了训练集中高频的词汇、实体和关系. 复制模式的词典 D 是动态变化的, 由每一次问答问题序列中所有的词构成. 知识库查询模式的词典 KB 同样是动态的, 包含每一次问答输入到模型中知识子图包含的实体. 3 个词典的交集部分 $V \cap D$, $V \cap KB$, $KB \cap D$ 体现了 3 种模式的合作关系, 非交集部分体现了 3 种模式的竞争关系. 空白部分的 UNK 是 3 个词典都不包括的陌生词汇.

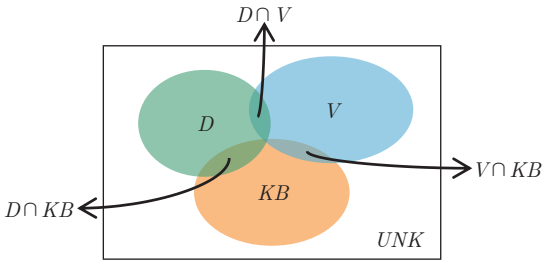


图 2　模型词典示意图

Fig. 2　The diagram of vocabulary

使用词向量作为基础语义单元的表达能力比使用字向量更强, 但使用字向量可以使模型 “理解”文本和知识库的时候可以在字的层面上考虑字符的相似性, 从而匹配可能性更高的知识. 并且字向量可以让所有陌生词汇有唯一的向量表征, 赋予了模型“理解”陌生词汇的能力. 所以本文使用字词向量结合的方式.

对问题序列 $X=[x_1, \cdots, x_{L_q}]$, 定义位置 j 的字词向量为 $\bar{\boldsymbol{x}}_j=[\boldsymbol{x}_j, \boldsymbol{u}_j]$, 其中 $\boldsymbol{x}_j$ 为序列位置 j 的词向量, L_q 为问题句子分词后的词数量. $\boldsymbol{u}_j$ 为位置 j 对应的字向量. 位置 j 的字向量表示为 $\boldsymbol{u}_j=[\boldsymbol{\chi}_1, \cdots, \boldsymbol{\chi}_{L_j}]$, 其中 L_j 为 x_j 包含的字符数. 输入中最长的问题序列长度为 $L_{\max}$, 每一个词最大包含的字符数 L_v. 当 $L_j<L_v$ 时需要在后面补齐差值的空字符 ⟨pad⟩, 当 $L_q<L_{\max}$ 时需要补齐 ⟨unk⟩ 词.

2.2　表示学习向量

为进一步提升模型匹配知识的能力, 本文利用知识表示学习方法 PTransE[24] 生成的实体、关系向量代替基础词典中相应词汇的词向量.

TransE[25] 模型将实体和关系在低维的空间里进行表达, 得到连续的向量, 将两个实体之间抽象的关系映射为两个向量之间的转换关系. TransE 模型只考虑了实体之间的直接关系, PtransE 模型在其基础上扩展, 将多步关系路径视为实体之间的连接.

PtransE 模型的优化目标函数主要分为两个部分: 一部分是实体之间关系约束; 另一部分是实体之间的路径约束.

对于一个三元组 (h, r, t) ($h, t\in E$, E 是实体集合, $r\in R$, R 为关系集合), 能量函数的定义如下: $E(h, r, t)=||\boldsymbol{h}+\boldsymbol{r}-\boldsymbol{t}||$, 来确保 $\boldsymbol{r}\approx\boldsymbol{t}-\boldsymbol{h}$ 成立. 对于实体对 (h, t) 之间存在的路径集合 $P(h, t)=\{p_1, \cdots, p_N\}$, 关系路径 $p=(r_1, \cdots, r_l)$ 表示 $h\xrightarrow{r_1}\cdots\xrightarrow{r_l}t$. 对路径三元组 (h, p, t), 能量函数定义如下: $E(h, p, t)=||\boldsymbol{p}-(\boldsymbol{t}-\boldsymbol{h})||=||\boldsymbol{p}-\boldsymbol{r}||=E(r, p)$. 路径 p 作为推理关系 r 的有效规则之一, 需要保证能量函数 $E(r, p)$ 可以得到尽可能低的分数. PtransE 模型目标函数为

$$L(S)=\sum_{(h, r, t)\in S}\left[L(h, r, t)+\frac{1}{Z}\sum_{p\in P(h, t)}R(p|h, t)L(p, r)\right] \tag{1}$$

其中, $L(h, r, t)$ 是对实体关系的约束, $L(p, r)$ 是路径关系约束. $R(p|h, t)$ 表示对实体对 (h, t) 路径 p 的置信度, PtransE 使用路径约束资源分配 (PCRA) 算法来衡量 $R(p|h, t)$. $L(h, r, t)$和 $L(p, r)$使用 Margin based 损失函数, 目标函数进一步表示为

$$L(h, r, t)=\sum_{(h', r', t')\in S^-}\left[\gamma+E(h, r, t)-E(h', r', t')\right]_+ \tag{2}$$

$$L(p, t)=\sum_{(h, r', t)\in S^-}\left[\gamma+E(p, t)-E(p, r')\right]_+ \tag{3}$$

其中, $[a]_+=\max(0, a)$, γ 是边界参数, S 表示知识库中的有效三元组, S^- 是负例, 对有效三元组随机替换其头和尾, 关系为: $S^-=\{(h', r, t)\}\cup\{(h, r', t)\}\cup\{(h, r, t')\}$.

通过表示学习方法对知识库建模, 将知识库中的实体、关系等语义单元表示为数值空间中的向量, 向量中的每一维数值表示该语义单元在 某一语义维度上的投影. 由于实体和关系的数值是根据整个知识库得到, 利用了整个知识库的特性, 从而包含更加全面的信息, 让模型中陌生词汇有了唯一的向量表征, 关系和实体向量有了更精确的语义表征.

2.3　编码器

问题的编码使用双向循环神经网络作为编码器. 输入端问题序列 $X=[x_1, \cdots, x_{L_q}]$, L_q 是问题序列的长度, 前向循环生成网络的隐藏层输出为 $\{\overrightarrow{\boldsymbol{h}}_1, \cdots, \overrightarrow{\boldsymbol{h}}_{L_q}\}$, 反向循环生成网络隐藏层输出为 $\{\overleftarrow{\boldsymbol{h}}_1, \cdots, \overleftarrow{\boldsymbol{h}}_{L_q}\}$, 编码器输出为前后向隐藏层状态的拼接: $\boldsymbol{h}_t=[\overrightarrow{\boldsymbol{h}}_t, \overleftarrow{\boldsymbol{h}}_t]$, 问题的记忆单元为 $\boldsymbol{M}_{\mathrm{Q}}=\{\boldsymbol{h}_1, \cdots, \boldsymbol{h}_{L_q}\}$, 储存编码器的所有隐藏状态. 问题的表征用所有隐藏状态的平均值来表示: $\boldsymbol{q}=\frac{1}{L_q}\sum_{j=1}^{L_q}\boldsymbol{h}_j$.

根据给定的主题词, 我们从相应的知识库中检索相关事实作为知识子图, 知识库中的事实表示为头、尾、关系的向量拼接[26]: $\boldsymbol{f}=[\boldsymbol{h}, \boldsymbol{t}, \boldsymbol{r}]$, 编码后得到知识库的记忆模块 $\boldsymbol{M}_{\mathrm{KB}}=\{\boldsymbol{f}_1, \cdots, \boldsymbol{f}_{L_{\mathrm{KB}}}\}$, L_{KB} 表示候选事实数量的最大值.

2.4　解码器

如图 1 所示, 模型使用单向循环生成网络 (Recurrent neural network, RNN) 作为解码器, 输出

答案序列: $Y=[y_1,\cdots,y_{L_a}]$, 其中 L_a 为长度. 模型结合了 CopyNet 框架和外部知识库, 解码器端按功能划分为: 生成模式、复制模式和 KB 查询模式. 复制模式负责将问题序列中的一些词汇复制到答案序列中; KB 查询模式负责查询 KB 中对应的知识指导答案生成; 生成模式负责生成连贯的自然语言串联另外两个部分. 模型在解码器端使用了全局覆盖机制来平衡模型 3 个模式的生成策略.

为进一步描述解码器, 本文将其工作过程按先后分为 3 个部分: 读取信息 (Read)、更新状态 (Update) 和输出预测 (Predict), 如图 3 所示.

2.4.1 信息读取

Read 部分是模型获取问题、知识库信息的过程. 在编码器一节介绍了模型通过双向 RNN 获取了问题的记忆模块 (Question memory) 记为 $\boldsymbol{M}_{\rm Q}$, 和知识库的记忆模块 (KB memory) 记为 $\boldsymbol{M}_{\rm KB}$. 模型复制模式通过读取 $\boldsymbol{M}_{\rm Q}$ 中的信息决定复制哪部分问题到答案中, 模型的 KB 查询模式通过读取 $\boldsymbol{M}_{\rm KB}$ 决定答案使用哪些事实作为知识依据.

2.4.2 状态更新

图 3 中的状态更新 (State update) 单元内部封装了 NMT (Neural machine translation)[27], NMT 的隐状态更新策略 $\boldsymbol{s}_t=f(\boldsymbol{s}_{t-1},y_{t-1},\boldsymbol{c}_t)$, 除了需要获知 $t-1$ 时刻解码器的隐状态 $\boldsymbol{s}_{t-1}$, 使用注意力机制生成的在时刻 t 问题序列的上下文向量 $\boldsymbol{c}_t$ 以及前一时刻的预测词 y_{t-1} 外, 在此基础上模型还需要获取上一个时刻源端和知识库的情况.

通过利用 $t-1$ 时刻源端和知识库的加权表征补充 y_{t-1} 缺失的信息, y_{t-1} 表示为: $[\boldsymbol{e}(y_{t-1}),\boldsymbol{r}_{q_{t-1}},\boldsymbol{r}_{kb_{t-1}}]$. 其中, $\boldsymbol{e}(y_{t-1})$ 是 y_{t-1} 对应字词向量. $\boldsymbol{r}_{q_{t-1}}$ 是问题的选择读取, 需要计算 $t-1$ 时刻 $\boldsymbol{M}_{\rm Q}$ 所有位置与 $\boldsymbol{s}_{t-1}$ 的相似度得分, 其中相似度得分计算函数 score$(\cdot)$ 使用向量点积计算公式, 再使用 softmax 函数获得上一时刻源端的注意力权重为

$$\alpha_{t-1j}=\frac{{\rm e}^{score(\boldsymbol{s}_{t-1},\boldsymbol{h}_j)}}{\sum\limits_{i=1}^{L_q}{\rm e}^{score(\boldsymbol{s}_{t-1},\boldsymbol{h}_i)}} \tag{4}$$

通过式 (5) 得到问题的加权表征为

$$\boldsymbol{r}_{q_{t-1}}=\sum_{j=1}^{L_q}\alpha_{t-1j}\boldsymbol{h}_j \tag{5}$$

对知识库的加权表征计算方式类似, $\boldsymbol{r}_{kb_{t-1}}$ 是知识库的选择读取, 需要计算 $t-1$ 时刻 $\boldsymbol{M}_{\rm KB}$ 每一个事实与 $\boldsymbol{s}_{t-1}$ 的相似度得分, 相似度得分计算函数 score$(\cdot)$ 同上, 再使用 softmax 函数计算注意力权重

$$\beta_{t-1j}=\frac{{\rm e}^{score(\boldsymbol{s}_{t-1},\boldsymbol{f}_j)}}{\sum\limits_{i=1}^{L_{\rm KB}}{\rm e}^{score(\boldsymbol{s}_{t-1},\boldsymbol{f}_i)}} \tag{6}$$

通过式 (7) 得到知识库的加权表征

$$\boldsymbol{r}_{{\rm kb}_{t-1}}=\sum_{j=1}^{L_{\rm KB}}\beta_{t-1j}\boldsymbol{f}_j \tag{7}$$

2.4.3 输出预测

MCQA 模型的目标端输出概率是由生成模式

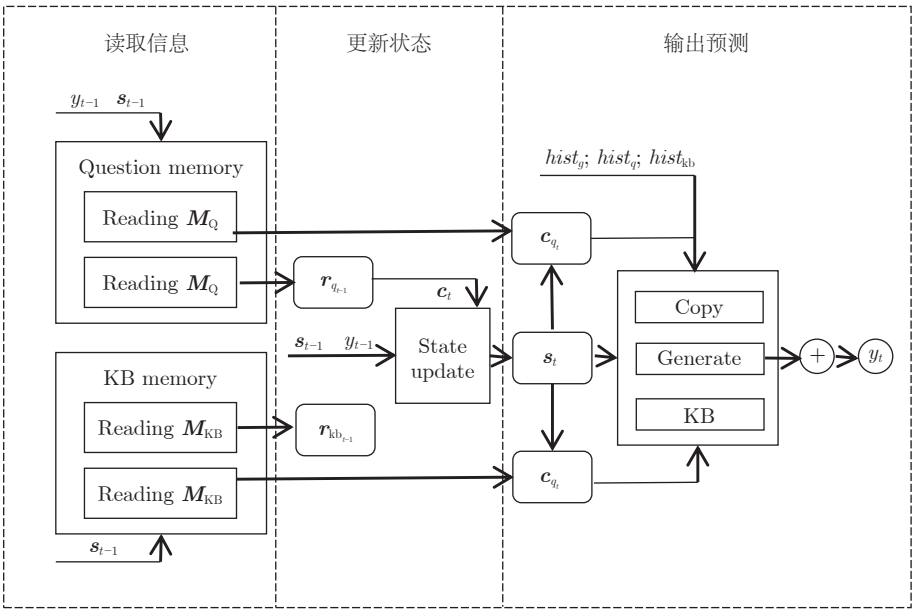


图 3 解码器工作机制示意图

Fig. 3 The diagram of working mechanism of decoder

$p(y_t, g|\cdot)$、复制模式 $p(y_t, c|\cdot)$、KB 查询模式 $p(y_t, kb|\cdot)$ 共同决定的, 3 个部分之间既有合作也有竞争, 得分最高的 y_t 是模型 t 时刻的输出, 输出的条件概率为

$$\begin{aligned} p(y_t|y_{t-1}, \boldsymbol{s}_t, \boldsymbol{M}_{\mathrm{Q}}, \boldsymbol{M}_{\mathrm{KB}}) = \\ p_{\mathrm{gen}}(y_t, g|y_{t-1}, \boldsymbol{s}_t, \boldsymbol{M}_{\mathrm{Q}}, \boldsymbol{M}_{\mathrm{KB}}, \boldsymbol{hist}_g) + \\ p_{\mathrm{cop}}(y_t, c|y_{t-1}, \boldsymbol{s}_t, \boldsymbol{M}_{\mathrm{Q}}, \boldsymbol{hist}_q, \boldsymbol{hist}_g) + \\ p_{\mathrm{kb}}(y_t, kb|y_{t-1}, \boldsymbol{s}_t, \boldsymbol{M}_{\mathrm{KB}}, \boldsymbol{hist}_{\mathrm{kb}}, \boldsymbol{hist}_g) \end{aligned} \tag{8}$$

模型通过加入 $\boldsymbol{hist}_q$ 和 $\boldsymbol{hist}_{\mathrm{kb}}$ 分别获取解码器在前 $t-1$ 时刻对问题序列的历史关注度总和以及对知识库中事实的历史关注度总和, 通过这些历史信息减小高关注度的部分受到的关注度, 缓解 Seq2Seq 模型中的过译问题.

$\boldsymbol{hist}_q$ 是 L_q 维的向量, 向量每一维记录问题序列每一个位置上的单词到目前的关注累积. 位置 k 词汇的总关注度为

$$hist_q^{(k)} = \sum_{\tau<t} \alpha_{\tau k} \tag{9}$$

其中, $\alpha_{\tau k}$ 是 τ 时刻第 k 个单词的关注度, 即复制模式下对每一个输出计算的得分.

$\boldsymbol{hist}_{\mathrm{kb}}$ 是 L_{kb} 维的向量, 它是解码器在前 $t-1$ 时刻对知识子图每一条事实的历史关注度总和, 向量每一维记录每一条事实到目前的关注累积. 第 k 条知识的总关注度为

$$hist_{\mathrm{kb}}^{(k)} = \sum_{\tau<t} \beta_{\tau k} \tag{10}$$

其中, $\beta_{\tau k}$ 是 τ 时刻第 k 条知识的关注度, 即 KB 查询模式下对知识子图中每一条事实计算的得分.

$\boldsymbol{hist}_g$ 是本模型加入的全局覆盖机制. 3 个模式在答案的生成过程中, 应该在总体上保持一种平衡. 参考人类回答问题的过程, 在简单的单轮知识对话中, 为保证答案的简洁和有效, 人类会拷贝问题中几个关键词语, 会从知识储备中选取 对应的知识, 然后用简单的话串联所有信息反馈回提问者, 即很少出现答案序列一直在由某一种模式生成, 所以模型需要一个变量来控制 3 个模式的切 换, 以选择是从基础词表生成单词, 还是从输入序列复制单词, 抑或是从知识库中选择事实构成答案. 当某个模式关注度足够高的时候, 提高其他模式的受关注的概率. 本模型通过 $\boldsymbol{hist}_g$ 实现 3 个模式的切换.

$\boldsymbol{hist}_g$ 是一个 3 维的向量, 它是解码器在前 $t-1$ 时刻对 3 种模式的历史关注度总和, 向量每一维记录每一种模式到目前的关注累积, 每一维度分别表示为: 生成模式历史关注度 $hist_g^{(\mathrm{gen})}$, 复制模式历史关注度 $hist_g^{(\mathrm{cop})}$, KB 查询模式历史关注度 $hist_g^{(\mathrm{kb})}$, 计算方法为

$$hist_g^{(\mathrm{gen})} = \sum_{\tau<t} \eta_{\tau \mathrm{gen}} \tag{11}$$

其中, $\eta_{\tau \mathrm{gen}}$ 是 τ 时刻模型对生成模式的关注度, 即 τ 时刻生成模式的条件概率.

$$hist_g^{(\mathrm{cop})} = \sum_{\tau<t} \eta_{\tau \mathrm{cop}} \tag{12}$$

其中, $\eta_{\tau \mathrm{cop}}$ 是 τ 时刻模型对复制模式的关注度, 即 τ 时刻复制模式的条件概率.

$$hist_g^{(\mathrm{kb})} = \sum_{\tau<t} \eta_{\tau \mathrm{kb}} \tag{13}$$

其中, $\eta_{\tau \mathrm{gen}}$ 是 τ 时刻模型对 KB 查询模式的关注度, 即 τ 时刻 KB 检索模式的条件概率.

模型利用式 (14) 计算全局平衡因子, 利用 t 时刻的状态 $\boldsymbol{s}_t$ 和全局覆盖向量 $\boldsymbol{hist}_g$ 来生成平衡因子 $\boldsymbol{\delta}$, 拼接 $\boldsymbol{s}_t$ 和 $\boldsymbol{hist}_g$ 经过参数为 $\boldsymbol{W}_g$ 的单层全连接网络层, $\boldsymbol{W}_g \in \mathbf{R}^{d_s + d_{\boldsymbol{hist}_g}}$, 通过 softmax 函数最终生成的平衡因子是一个维度为 3 的向量, 向量的每一维依次代表: 生成模式、复制模式和 KB 查询模式的平衡系数, 用来平衡 3 个模式的得分.

$$\boldsymbol{\delta} = [\delta_{\mathrm{gen}}, \delta_{\mathrm{cop}}, \delta_{\mathrm{kb}}] = \mathrm{softmax}(\boldsymbol{W}_g \times [\boldsymbol{s}_t, \boldsymbol{hist}_g] + b_g) \tag{14}$$

3 个模式通过全局覆盖机制来平衡得到最终的得分. 然后使用 softmax 函数得到生成模式、复制模式、KB 查询模式 3 个部分的条件概率, 在得分函数前乘上各自对应的平衡因子. 复制和 KB 查询两种模式与生成模式稍有不同, 以复制模式为例, 同一个词可能在问题序列中的不同位置出现多次, 所以计算该词时要综合每个位置的得分. 每个模式的条件概率分布计算方法为

$$p_{\mathrm{gen}}(y_t, g|\cdot) = \frac{1}{Z} \mathrm{e}^{\delta_{\mathrm{gen}} \times score_{\mathrm{gen}}(y_t)} \tag{15}$$

$$p_{\mathrm{cop}}(y_t, c|\cdot) = \frac{1}{Z} \sum_{j:Q_j = y_t} \mathrm{e}^{\delta_{\mathrm{cop}} \times score_{\mathrm{cop}}(y_t)} \tag{16}$$

$$p_{kb}(y_t, kb|\cdot) = \frac{1}{Z} \sum_{j:KB_j = y_t} \mathrm{e}^{\delta_{\mathrm{kb}} \times score_{\mathrm{kb}}(y_t)} \tag{17}$$

其中, Z 是 3 个模式经过 softmax 函数计算的归一化项 $Z = \mathrm{e}^{\delta_{\mathrm{gen}} \times score_{\mathrm{gen}}(y)} + \sum_{j:Q_j=y} \mathrm{e}^{\delta_{\mathrm{cop}} \times score_{\mathrm{cop}}(y)} + \sum_{j:KB_j=y} \mathrm{e}^{\delta_{\mathrm{kb}} * score_{\mathrm{kb}}(y)}$. $score_{\mathrm{gen}}$, $score_{\mathrm{cop}}$和 $score_{\mathrm{kb}}$分别是 3 种模式从词表选择语义单元的评分函数, 它们的计算过程如下:

1) 生成模式

此模式是解码端的基本模式, 用于生成基础词典中的词. 如式 (18) 所示, $\boldsymbol{s}_t$ 通过张量 $\boldsymbol{W}_{\mathrm{gen}}$ 得到词

典中每一个词的得分, $\boldsymbol{v}_i$ 是 v_i 的独热 (one-hot) 表示, 通过点积获取 v_i 的得分. $\boldsymbol{W}_{\text{gen}} \in \mathbf{R}^{(d_h+d_i+d_f)\times d_o}$, 其中, d_h, d_i 和 d_f 分别表示 RNN 隐藏层向量、输入词向量和知识库事实向量的维数.

在本模型中, $\boldsymbol{s}_t$ 后拼接了 $\boldsymbol{c}_{q_t}$ 和 $\boldsymbol{c}_{kb_t}$ 两个向量, 意义在于通过此时的问题的加权表征和知识库的加权表征来指导生成模式的预测. 例如知识库中的知识为 (A, 性别, 男), 但模型在生成答案时需要的不是直接将性别属性拷贝, 而是需要用 "男"这条知识来指导生成模式生成 "他"这个人称代词, 同理在问题序列中出现如 "中国"等国籍信息, 模型需要将 "中国"变换为 "中国人". $\boldsymbol{c}_{q_t}$ 和 $\boldsymbol{c}_{kb_t}$ 是模型使用 Attention 机制在当前时刻的选择读取, 计算类似于 $\boldsymbol{r}_{q_t}$, $\boldsymbol{r}_{kb_t}$.

$$score_{\text{gen}}(y_t = v_i) = \boldsymbol{v}_i^{\text{T}} \boldsymbol{W}_{\text{gen}}[\boldsymbol{s}_t, \boldsymbol{c}_{q_t}, \boldsymbol{c}_{kb_t}] \tag{18}$$

2) 复制模式

此模式负责将问题序列中部分词汇原封不动地复制到答案序列, 复制模式会构建一个新的词典, 其中保存没有在基础词典中出现的词, 模型会预测需要复制问题序列的第几个词. 如式 (19) 所示, x_j 是问题序列第 j 个词, $\boldsymbol{h}_j$ 是 x_j 的解码器输出, 本模式将 $\boldsymbol{hist}_q$ 视为 t 时刻的解码状态的一部分, $\boldsymbol{s}_t$ 和 $\boldsymbol{hist}_q$ 可以分别理解为 t 时刻模型的语义状态和复制模式的历史状态, 所以将二者进行拼接, 通过张量 $\boldsymbol{W}_c$ 和 tanh 激活函数计算与 $\boldsymbol{h}_j$ 的相似度得分, $\boldsymbol{W}_c \in \mathbf{R}^{(d_s+d_{L_q})\times d_h}$. $score_{\text{cop}}(y_t = x_j)$ 越高预测输出为 x_j 的概率越大.

$$score_{\text{cop}}(y_t = x_j) = \tanh([\boldsymbol{s}_t, \boldsymbol{hist}_q]^{\text{T}} \times \boldsymbol{W}_c)\boldsymbol{h}_j \tag{19}$$

3) KB 查询模式

此模式负责查询知识库, 将最匹配的事实填入到答案中, 本质上是一个应用于知识库的 CopyNet, 该模式同样会构建一个词典, 其中保存知识库中出现的 OOV 词, 模型会预测需要复制知识库中的第几个三元组的尾实体. 如式 (20) 所示, f_k 代表知识库中第 k 个三元组的尾实体, $\boldsymbol{f}_k$ 是三元组的向量表征. $\boldsymbol{s}_t$ 和 $\boldsymbol{hist}_{\text{kb}}$ 分别代表语义状态和 KB 查询的历史状态, 除此之外本模型额外加入问题序列的整体表征 $\boldsymbol{q}$, 直观上可以理解为在知识库的查询过程中要同时把握局部和全局的问题语义信息, 式 (20) 中将三者进行拼接作为最终的状态表征, 通过张量 $\boldsymbol{W}_{\text{kb}}$ 和 tanh 激活函数计算与 $\boldsymbol{f}_k$ 的相似度得分, $\boldsymbol{W}_{\text{kb}} \in \mathbf{R}^{(d_s+d_h+d_{L_{\text{KB}}})\times d_f}$.

$$score_{\text{kb}}(y_t = f_k) = \tanh([\boldsymbol{s}_t, \boldsymbol{q}, \boldsymbol{hist}_{\text{kb}}]^{\text{T}} \times \boldsymbol{W}_{\text{kb}})\boldsymbol{f}_k \tag{20}$$

2.5 模型优化

本文提出的 MCQA 模型使用端到端的反向传播模式进行优化, 给定一个 Batch 大小为 N 的训练数据: 问题集合 $\{X\}_N$、答案集合 $\{Y\}_N$ 和知识集合 $\{Z\}_N$. 本文使用最大似然函数作为优化目标, 损失函数使用负对数似然函数如式 (21) 所示, 使用梯度下降来优化损失函数. 因为 3 种模式共享相同的 softmax 分类器的预测目标词, 所以它们可以通过最大化似然函数的方式来彼此协调学习.

$$L = -\frac{1}{N}\sum_{k=1}^{N}\sum_{t=1}^{L_q} \log\left(p(y_t^{(k)}|y_{<t}^{(k)}, X^{(k)}, Z^{(k)}\right) \tag{21}$$

3 结合知识推理的模糊问答

本文使用表示学习方法 PTransE 对知识库建模, 应用知识推理方法对知识库进行补全, 通过数据共享的方式使问答模型 MCQA 具有模糊问答的能力.

3.1 知识推理

知识推理是知识库补全的过程, 在知识图谱中直观地表示为: 判断两个实体 (结点) 间是否可能存在着某种已知关系 (边), 如果置信度达到一定阈值便可以在两个实体间加入推理出的关系.

具体来说, 使用 PTransE 方法对知识库建模后得到的实体和关系向量来完成知识库补全任务. 已知三元组其中的两部分, 用来对第三部分进行预测. 对三元组 (h, r, t), 定义如下打分函数来计算候选项的得分. 函数定义为[24]

$$Score(h, r, t) = G(h, r, t) + G(t, r^{-1}, h) \tag{22}$$

其中, $G(h, r, t)$ 包含两部分: 一部分是关系损失 $||\boldsymbol{h}+\boldsymbol{r}-\boldsymbol{t}||$; 另一部分是路径损失 $R(p|h, t)||\boldsymbol{p}-\boldsymbol{r}||$. 综合两部分损失越低, 该备选三元组越有可能成为最终预测结果. 关系中的 -1 代表逆关系, 对三元组 (h, r, t) 调换头尾形成的三元组为 (h, r^{-1}, t), 关系 r^{-1} 是关系 r 的逆关系. 路径 p 的可靠性与给定 r 的推理强度有关, 该强度可从训练数据中量化为 $Pr(r|P) = Pr(r, p)/Pr(p)$. 最终得到 $G(h, r, t)$ 的计算式为

$$G(h, r, t) = ||\boldsymbol{h}+\boldsymbol{r}-\boldsymbol{t}|| + \frac{1}{Z}\sum_{p\in P(h,t)} Pr(r|P)R(p|h, t)||\boldsymbol{p}-\boldsymbol{r}|| \tag{23}$$

但是基于表示学习的推理方式不同于传统的基

于规则的推理机, 推理的结果即使达到很高置信度, 也是不确定的, 或是由现有知识不能判定正误, 因此本文认为应当在知识库补全的过程中区分原有知识和推理知识.

通过知识推理得出 e_i 和 e_j 之间存在关系 r 时 $(e_i \in E, e_j \in E)$, 在关系 r 中加入模糊词汇构成推理关系 r', 将推理知识 (e_i, r', e_j) 加入知识库.

在问答模型中, 使用知识表示学习向量得到的关系向量 $\boldsymbol{r}$ 与推理关系 $\boldsymbol{r}'$ 是共享关系向量的, 考虑两个关系的异同, 引入一个可训练的正则化偏置项 b

$$\boldsymbol{r}' = \boldsymbol{r} + \boldsymbol{b} \tag{24}$$

其中, $\boldsymbol{r}$ 是 r 的向量表示, $\boldsymbol{r}'$ 是 r' 的向量表示, 通过上述方法对数据集中的推理关系初始化.

3.2　训练数据

由于推理关系的不确定性, 模糊回答需要在不确定的答句中适当加入 “可能”、“应该”等推断词. 语料要求比较独特, 本文通过规则生成的方式在现有数据集的基础上构建新数据集. 根据推理关系 r' 以及主题词 e, 在现有数据集中寻找针对关系 r 问询的问答对, 替换问答对中的主题词, 以及答案中的尾实体信息, 并通过句法分析工具解析句子成分, 在特定部分加入推断词, 并将生成的问答对以及三元组加入新构建的数据集中.

在数据预处理阶段, 通过遍历找到实体与实体间可能存在路径和路径长度信息, 将关系以及实体之间的路径存储进行训练.

使用补全后的知识库作为知识问答模块的外部知识库. 由于加入了推理关系的概念, 模型可以有效地区分出哪些知识是原始知识, 哪些是推理知识, 能够对推理得到的知识进行模糊回答.

4　实验结果与分析

4.1　实验数据集

本文共使用 3 个数据集: SimpleQuestion 单关系知识问答数据集、生日问答限定领域数据集和社区问答开放领域数据集, 数据集详细信息如表 1 所示.

表 1　问答数据集规模

Table 1　The size of QA datasets

数据集	问答对数量	关系数量
SimpleQuestions	101 754	1 631
生日问答数据集	239 922	5
社区问答数据集	505 021	4 011

1) SimpleQuestions 数据集: 经典的 KBQA 英文数据集, 语料包括: 问题, 答案所依赖的三元组. 数据集是单一关系 (Single relation) 数据集, 每一个问答对都只依赖于一个事实. 数据集选自 freebase, 本文除该数据集以外使用 FB5M 和 FB2M 作为外部知识库. 数据集按照 7 : 1 : 2 的比例划分为训练集、验证集和测试集.

2) 生日问答数据集: 中文限定领域的生成式 KBQA 数据集, 由 CoreQA[6] 提出, 数据集是使用模板生成得生日的问答语料. 数据集的答案依赖多条事实. 数据集按照 9 : 1 的比例划分为训练集和测试集.

3) 社区问答数据集: 中文开放领域的生成式 KBQA 数据集. 由 CoreQA[6] 提出, 数据集包含了多个中文问答社区的问答语料. 语料包括: 问题、答案以及答案所依赖的多条事实, 语料规模庞大且涉及领域广阔. 数据集按照 9 : 1 的比例划分为训练集和测试集.

在 3 个数据集上对使用字词向量的模型 MCQA (WE, CE), 将词向量中的实体和关系向量替换为知识表示学习生成向量的模型 MCQA (TE, CE) 分别进行了实验.

4.2　单关系知识库问答任务

1) 任务描述

在单一关系的 SimpleQuestions 数据集中测试模型查找正确知识的能力. 任务要求模型在给定主题词的情况下在知识库中寻找答案, 本质上是在主题词对应的知识子图中识别正确的关系. 本文提出的模型是面向于生成式问答任务的, 但同样可以完成只返回正确实体的 KBQA 任务, 之前很多学者提出的模型在此已经达到了很好的效果. 实验证明, 本文提出的模型获得了优于 Baseline 以及多个近年提出的 KBQA 模型.

实验评判指标使用预测答案的准确率, 即预测正确的测试样本数除以总测试样本数, 准确率越高说明模型查找正确知识的能力越强.

2) 实验设置

本文实验在经过实体链接[13, 28] 后的数据集上进行, 经过统计, 在 FM5M 中 SimpleQuestion 的主题词的知识子图规模集中在 1 到 50 条三元组, 所以本次实验知识库最大长度 L_{kb} 设为 50, 不足 50 条的样本随机填入无关三元组或由 UNK 组成的空三元组. 为保证公平性与对比实验保持一致, 本文模型框架中 RNN 网络使用长短期记忆网络 (Long short-term memory, LSTM), 隐藏层大小设置为 256 维, 词向量维度为 256 维, 字向量维度设置为 2 维, 词的最大长度 L_c 设置为 8, L_{kb} 设置为 50, 使

用 Adam 作为梯度下降策略.

3) 结果分析

本文提出的 MCQA 模型在 SimpleQuestions 数据集的表现如表 2 所示, 此次实验使用文献 [13] 中的双向卷积神经网络 (Bi-directional CNN, BiCNN) 模型作为 Baseline, AMPCNN (Attentive max-pooling CNN)[29] 是加入字符向量和注意力机制的 CNN 模型, 文献 [30] 提出的 HR-BiLSTM (Hierarchical residual bi-directional LSTM) 模型在句子和关系两个级别分析语义, 准确率达到了 93%.

表 2 SimpleQuestion 数据集实验结果
Table 2 The experimental results of SimpleQuestion datasets

方法	准确率 (%)
BiCNN[13]	90.0
AMPCNN[29]	91.3
HR-BiLSTM[30]	93.3
CoreQA	92.8
MCQA (WE, CE)	93.8
MCQA (TE, CE)	**94.3**

MCQA(WE, CE) 是本文使用了字词向量的模型, 可以看到准确率达到了 93.8%, 超越了 HR-BiLSTM 模型, 进一步将词向量中的实体和关系向量替换为知识表示学习生成的向量 MCQA (TE, CE) 模型, 准确率进一步提高到 94.3%. CoreQA 与 MCQA 模型同为生成式问答模型, 可以看出, 字词向量和知识表示学习结果的加入明显提升了 MCQA 模型在知识库中查找正确知识的能力.

当前在该数据集上表现最好的 KBQA 模型是在 2018 年提出的多任务学习的模型[31], 达到了 95.7% 的准确率, 虽然本文提出的模型没能超过该模型, 但由于本模型是面向生成式 KBQA 任务的, 只需要验证其具有查询正确知识的能力即可. 实验证明, 模型可以在提供的规模为 50 条事实中选取出正确的事实, 加入字词向量和知识表示学习结果能进一步提高模型效果, 并能达到 94% 以上的准确率.

4.3 限定领域知识问答任务

1) 任务描述

在生日数据集上测试模型在限定领域中寻找正确知识的能力和生成答案的表述能力. 生日数据集是模拟现实中人类对某人生日的问答来生成的数据集, 本文通过对生成答案的年、月、日、性别的正确率进行统计来验证模型查询知识库的准确率, 使用正则表达式匹配的方式来判断模型是否生成了流畅连贯的高质量回答.

2) 实验设置

模型及相应的对比实验中所有的 RNN 使用 GRU (Gate recurrent unit) 网络, 隐藏层大小设置为 500 维, 词向量维度为 256 维, 字向量维度设置为 4 维, 词的最大长度 L_c 设置为 8, $L_{\rm kb}$ 设置为 5, 使用 Adam 作为梯度下降策略. 本次实验使用了问答语料对和知识库中的关系来生成基础词典.

本文通过正则表达式匹配的方法提取出测试答案的人称代词、年、月、日信息, 准确率分别表示为 P_g, P_y, P_m, P_d. 为评判生成答案的语言表述能力, 本文生成了 18 个语法正则模板, 评判标准 P_r 是符合模板规律的测试样本数除以测试总样本数, P_r 越高说明模型语言表述能力越强.

3) 结果分析

模型在生日数据集上表现如表 3 所示, "—"代表该项没有数据或正确率小于 10%. 前三组实验的模型没有知识库参与, 分别是基础 Seq2Seq 框架、神经翻译模型和复制网络模型根据问答对语料进行训练得到的结果, 虽然模型没有查找知识库的能力, 但是它们生成的答案的表述能力可以作为生成答案质量的 Baseline.

表 3 生日数据集实验结果 (%)
Table 3 The experimental results of birthday datasets (%)

方法	P_g	P_y	P_m	P_d	P_r
Seq2Seq	67.3	—	23.4	—	37.2
NMT	71.6	—	27.1	—	54.7
CopyNet	75.2	—	—	—	71.9
GenQA (本文)	73.4	63.2	65.8	77.1	62.6
CoreQA	75.6	84.8	93.4	81	80.3
MCQA (WE, CE)	**89.8**	89.1	98.4	93.2	84.1
MCQA (TE, CE)	88.6	**89.4**	**98.7**	**93.6**	**84.6**

GenQA 和 CoreQA 是经典的生成式 KBQA 模型, 由于 GenQA 原始模型是面向单个知识进行回答的, 所以我们将 GenQA 模型稍作修改, 使其可以根据多个知识回答问题, 效果如 GenQA (本文) 所示. CoreQA 模型与本文的 MCQA 模型类似, 是结合了 GenQA 与复制网络的模型, 也是本文的主要对比实验.

由于本文提出的 MCQA 模型与 CoreQA 模型相比加入了全局覆盖向量记录 3 个模式的历史关注度, 防止出现某个模式训练过拟合而其他模式训练不充分的情况, 当某个模式关注度过高时提升其他模式的受关注概率, 因此 P_r 一项提高了近 4%, 达到了远高于 Baseline 的效果.

性别预测一项, 由于实验将预测出实体名称视为预测正确, 所以所有模型普遍准确率较高, 但是事实上为了检验模型处理知识库中 OOV 词的能力, 实验中所有模型的基础词典中我们故意忽略了“男”、“女”两个词, 知识库的性别信息无法识别, 所有的对比模型包括 MCQA 的性别预测基本等于随机. MCQA(WE, CE) 加入了字词向量表示, 增强了模型的 “理解”能力, 所以在性别预测一项的准确率有大幅度的提升. 在年月日三项的预测中, 加入了字词向量后的 MCQA 进一步得到了提升, 因为关系词汇中的 “年”、“月”、“日”等字符表征能够匹配到问题序列中的相同字符, 提高了模型的查询能力. 加入表示学习的实体关系向量后模型提升较小, 是因为该数据集的知识图谱比较稀疏, 仅有 5 种关系, 能够通过表示学习得到的隐式信息有限.

4.4　开放领域的社区问答任务

1) 任务描述

在大规模社区问答数据上测试模型生成答案的质量和拟合大规模数据的能力. 相比生日数据集该数据集涉及领域广泛, 具有更多种类的关系和大规模的实体, 会产生更多的 OOV 词. 该实验同时体现了模型在互联网应用的现实意义, 如果模型可以生成正确的答案, 而且答案表述与人类似, 那么深度学习模型便可以代替人工回答部分简单问题.

2) 实验设置

模型及对比实验中所有的 RNN 使用 GRU 网络, 隐藏层大小设置为 500 维, 词向量维度为 256 维, 字向量维度设置为 6 维, 词的最大长度 L_c 设置为 8, L_{kb} 设置为 10, 不足的部分使用无关三元组或空三元组补齐, 使用 Adam 作为梯度下降策略. 问答对语料经过命名实体识别处理, 分词工具使用 jieba 分词.

由于开放领域的社区问答生成的答案构成比较复杂, 没有统一的计算模型效果的方法, 所以本实验结果采用人工检验的方式, 每次随机选取 100 条来进行检验, 共检验 3 次取平均值, 从流畅性、一致性、正确性检验答案的质量, 分别统计符合 3 个指标的测试样本数量除以总样本数计算符合 3 个指标的得分概率, 概率越高模型在该指标上表现越好. 流畅性是指答案是否符合句法语法结构, 一致性是指答案是否与问题的方向保持一致, 正确性指是否答案中包含正确的知识. 需要注意的是数据集中会存在质量很差的样本如知识库与问题无关等情况, 所以在验证时低质量样本不计入统计中.

3) 结果分析

如表 4 所示, 实验中 CopyNet 模型是流畅性和一致性指标的 Baseline, GenQA 模型是所有 3 个指标的 Baseline. 由于实验使用开放领域的社区问答语料, OOV 问题体现得更加明显, 在对比实验生成的答案中出现了大量的 ⟨unk⟩ 词汇. 与对比实验 CoreQA 相比, 模型在正确性、流畅性和一致性 3 个指标上均超越了 CoreQA. MCQA(WE, CE) 加入字词向量提升了模型 “理解”问题和知识库的能力, 模型能明确问题的方向并找到正确答案, 在一致性和正确率方面有明显提升, 模型在加入知识表示学习结果后, 缓解 OOV 带来的影响, 让陌生词汇有了唯一表征, 利用了整个知识库的信息, 在正确性上得到了进一步提升. 由于 MCQA 的全局覆盖机制, 能够平衡模型 3 个模式的生成策略, 使得语言逻辑更加清晰, 语言流畅性更高, 因此在流畅性指标上较对比实验有较大提升.

表 4　社区问答实验结果 (%)

Table 4　The experimental results of community QA datasets (%)

方法	正确性	流畅性	一致性
CopyNet	—	19.4	21.3
GenQA (本文)	24.3	38.3	24.1
CoreQA	49.3	51.8	62.5
MCQA (WE, CE)	52.3	55.8	**65.2**
MCQA (TE, CE)	**54.1**	**56.3**	65.0

在图 4 和图 5 中我们列举了一些实验的例子, 图中预测答案里普通字体的部分由生成模式产生, 下划线部分由复制模式产生, 加粗部分由 KB 查询模式产生, 斜体部分是由 KB 知识指导生成模式产生的词汇.

在图 4 中, 我们展示了 MCQA 与对比实验 CoreQA 生成的一些错误案例的对比. 在问句 1 中, 对比实验 CoreQA 生成的答案多了一个 “的”, 原因在于模型可能过拟合于一种语言模式. 而 MCQA 加入全局覆盖机制后能够平衡 3 种模式的生成策略, 没有出现这种情况. 在问句 2 中, 对比实验 CoreQA 在答案中生成重复的事实, 原因可能是模型过拟合于 KB 查询模式并且不同时刻事实 “周树人”都有很高的关注度, 在答案中连续生成多个相同的词. 在 MCQA 中因为加入了全局覆盖机制控制模式切换没有出现这种错误. 在问句 3 中, 对比实验 CoreQA 生成了无意义重复答案, 可能因为模型没有从知识库中检索出正确的知识, 也可能因为模型陷入生成模式无法跳出. 而 MCQA 通过加入字词向量和表示学习结果增强了模型的 “理解”能力, 提高了模型检索知识的准确率, 并且加入全局覆盖机制控制模型在不同模式的切换, 能够对答案

ID	问题	MCQA	CORE
1	很喜欢，西游记的作者是谁？谢谢	西游记作者是**吴承恩**	西游记作者是**吴承恩**的
2	鲁迅原名是什么？	原名**周树人**	原名**周树人周树人周树人**
3	我需要周瑜的详细资料，急用！！！	**周瑜**字**公瑾**	是是是

图 4 MCQA (TE, CE) 与 CoreQA 答案对比样例

Fig. 4 The comparison of MCQA (TE, CE) and CoreQA sample outputs

ID	问题	预测答案	标准答案
1	三国时期的吕布到底是怎样的历史？	是三国时期的**武将**	三国第一武将
2	金庸为什么叫“金庸”？	金庸原名**“查良镛”**？	可能是金庸比查良镛听起来顺耳吧
3	鲁迅名什么又名什么？	**周树人**，字***豫才***	周树人
4	海绵宝宝走了，派大星会孤独么？	是**章鱼哥**的<unk>	他还有章鱼哥
5	很喜欢，西游记的作者是谁？谢谢	西游记的作者是**吴承恩**	吴承恩
6	大银行家摩根活了多大?	**1837年4月17日—1913年3月31日**	1837年4月17日—1913年3月31日自己算
7	泡泡堂是哪个公司开发的？	泡泡堂是**盛大网络**公司开发的	简单，盛大网络
8	诸葛亮和郭嘉，大家更喜欢哪一个?	更喜欢**孔明**咯～	孔明咯～
9	孔子是什么朝代人？	**春秋**时期**鲁国人**	春秋战国鲁国人

图 5 社区问答样例

Fig. 5 The sample outputsof community QA

生成较完整流畅的正确句子.

图 5 将 MCQA 生成的答案和标准答案进行比对, 整体上答案的语义结构比较清晰, 也能较好地与问题的方向保持一致. 但生成的答案中依旧存在待解决的问题, 我们对错误案例进行分析, 如图 5 的 ID4 例子所示, 结果不可避免地出现了 ⟨unk⟩ 单词, 也证明了本文提出的针对 OOV 词的解决方案, 只能缓解 OOV 词为模型带来的影响, 而不能彻底解决 OOV 问题. ID8 的答案没有明显问题, 但结合知识库看, 知识库只提供了 (诸葛亮, 别名, 孔明) 的知识, 训练集中也没有类似的训练数据, 因此在没有知识指导的情况下, 模型会通过学到的语言模型随机生成答案.

4.5 模糊问答任务

1) 任务描述

在知识问答模型 MCQA 的基础上应用 PTransE 方法, 推理知识库隐含的知识, 对推理出的知识进行模糊回答, 并保证模型在回答原有知识和现有知识的准确率, 以及生成答案的质量.

数据集构建: 任务使用生日数据集作为原始数据, 为构建符合要求的新数据集, 本文在知识库原有的 “性别”、“出生年”、“出生月”、“出生日”的基础上加入了 “同年生”、“同月生”、“同日生” 3 个关系, 重新生成 40 000 个主题词及其部分生日相关三元组, 缺失的生日信息通过新加入的 3 个关系对应到原有知识中, 例如新主题词缺失 “出生日”的信息, 通过 “同日生”关系将其联系到原有知识中, 最后使用模板对新加入主题词生成问答对.

任务流程: 对新加入的主题词的 “出生年”、“出生月”、“出生日”三种关系进行推理, 对应的推理关系为 “推断出生年”、“推断出生月”、“推断出生日”. 将推理知识加入知识库中, 如图 6 所示, 实体 ⟨Ent_111018⟩ 缺失 “出生年”知识, 通过与 ⟨Ent_26747⟩ 实体 “同年生”的知识推理出可能的出生年为 “1978”, 并将推理出的三元组 (Ent_111018, 推断出生年, 1978) 加入到知识库中. 重新划分训练集和测试集, 使用 MCQA 模型进行问答任务.

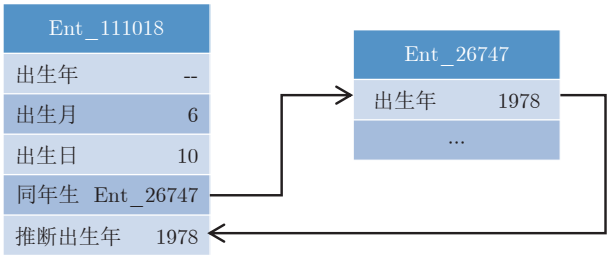


图 6 知识补全示意图

Fig. 6 The diagram of knowledge base completion

ID	正确知识	问题	模糊问答结果
1	1945.7.27	Ent_80329 的出生时间是什么呢?	他**可能**是 1945 年 7 月 27 日出生的
2	1995.8.8	Ent_40076 的生日是什么时候?	她的生日是 8 月 8 日
3	***1927***.2.12	Ent_118514 是什么时候出生的?	Ent_118514 ***大概率***是 1927 年 2 月 12 日出生
4	***1991.6.22***	Ent_98874 是多会出生的?	她***应该***是 1991 年 6 月 22 日出生
5	1907.2.23	Ent_107981 的出生年月日是什么?	他的出生年月日是 1907 年 2 月 23

图 7　模糊问答样例

Fig. 7　The sample outputs of ambiguously QA

2) 实验设置

知识推理模型采用参数如下: 实验使用 2 步 (2-step) 的所有路径作为特征路径, 梯度下降的学习率 α、批量 b、向量维度 d、hinge 损失函数的 γ、协调参数 θ. 实验使用如下参数取值: $\alpha = 0.001$, $b = 100$, $d = 100$, $\gamma = 1$, $\theta = 1.7$, 训练迭代次数为 200 次.

知识问答模型采用参数如下: 隐藏层大小设置为 500 维, 词向量维度为 256 维, 字向量维度设置为 4 维, 词的最大长度 L_c 设置为 8, L_{kb} 设置为 10, 使用 Adam 作为梯度下降策略.

3) 结果分析

如表 5 所示, 模糊问答任务的知识推理结果, 实验统计了所有新主题词出生年、月、日三个关系的准确率 P_y, P_m, P_d, 推理的效果达到了准确率 90% 以上.

表 5　模糊问答推理结果 (%)

Table 5　The prediction results of ambiguously QA (%)

方法	P_y	P_m	P_d
PTransE	93.2	97.4	95.0

如表 6 所示, 模糊问答的结果, $F1_t$ 是评判模型是否能正确判断, 是否需要采用模糊回答的 F1 值, $F1_t$ 达到了 87.7%, 说明 MCQA 模型已经初步具备了模糊回答的能力. P_g, P_y, P_m, P_d 统计了模型根据原始数据和推理数据回答的答案准确率, 也都达到了较好的效果. P_r 也达到了一个较高的水平, 意味着加入推理词后答案依旧具有很好的可读性.

表 6　模糊问答结果 (%)

Table 6　The results of ambiguously QA (%)

方法	$F1_t$	P_y	P_m	P_d	P_r
MCQA (WE, CE)	87.7	78.1	88.2	90.8	80.9

模糊回答举例如图 7 所示, 正确知识一列中普通字体的部分是知识库中原有知识, 斜体的部分是经过知识推理后得到的推理知识, 可以看到模型使用推理知识后会使用模糊回答的方式在答案适当位置加入推断词, 而只使用原有知识的答案中会直接给出肯定的答案.

5　结束语

本文提出了一种基于表示学习与全局覆盖机制的生成式知识问答模型, 针对现有的生成式问答任务中遇到的 OOV 问题, 本文提出引入知识表示学习方法生成的实体、关系向量代替基础词典中相应词汇的词向量, 提高模型识别陌生词汇的能力, 提高模型准确率. 针对模型在使用 3 种预测模式联合生成答案时模式混乱的问题, 本文提出了全局覆盖这一机制, 提高模型的语义连贯性, 减少由预测模式混乱导致的重复输出的问题. 另外本文还将知识表示学习结果应用到问答模型中, 使模型初步具有区分原始知识和推理知识的能力, 能对推理知识进行模糊回答. 模型在多个数据集上的实验都取得了很好的效果, 表明模型能够有效提高生成答案的准确率和答案的语义流畅度. 最后, 知识推理和问答模型的结合使模型对推理知识的模糊回答取得了一定的效果.

未来我们将继续研究将知识表示学习和生成式问答进一步深入结合, 在问答模型中嵌入知识推理模型, 进行多任务联合学习, 改善学习性能. 还会尝试在问答模型的训练过程中加入课程学习的思想, 先易后难的训练模型, 使得含有噪声的数据中所有有价值的信息都可以得到充分利用.

References

1　Vanessa L, Victoria U, Marta S, Enrico M. Is question answering fit for the semantic web? A survey. *Semantic Web*, 2011, **2**(2): 125−155

2　Sydorova A, Poerner N, Roth B. Interpretable question answering on knowledge bases and text. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence, Italy: ACL, 2019. 4943−4951

3 Cho K, Merriënboer B V, Gulcehre C, Bahdanau D, Bougares F, Schwenk H, Bengio Y. Learning phrase representations using rnn encoder-decoder for statistical machine translation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. Doha, Qatar: Association for Computational Linguistics, 2014. 1724−1734

4 Gu J T, Lu Z D, Li H, Li V O K. Incorporating copying mechanism in sequence-to-sequence learning. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. Berlin, Germany: the Association for Computational Linguistics, 2016. 1631−1640

5 Gulcehre C, Ahn S, Nallapati R, Zhou B W, Bengio Y. Pointing the unknown words. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. Berlin, Germany: the Association for Computational Linguistics, 2016. 140−149

6 He S Z, Liu C, Liu K, Zhao J. Generating natural answers by incorporating copying and retrieving mechanisms in sequence-to-sequence learning. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. Vancouver, Canada: the Association for Computational Linguistics, 2017. 199−208

7 Liu Kang, Zhang Yuan-Zhe, Ji Guo-Liang, Lai Si-Wei, Zhao Jun. Representation learning for question answering over knowledge base: An overview. *Acta Automatica Sinica*, 2016, **42**(6): 807−818
(刘康, 张元哲, 纪国良, 来斯惟, 赵军. 基于表示学习的知识库问答研究进展与展望. 自动化学报, 2016, **42**(6): 807−818)

8 Tang X, Chen L, Cui J, Wei B G. Knowledge representation learning with entity descriptions, hierarchical types, and textual relations. *Information Processing and Management*, 2019, **56**(3): 809−822

9 Mikolov T, Chen K, Corrado G, Dean J. Efficient estimation of word representations in vector space. arXiv: 1301.3781, 2013.

10 Unger C, Freitas A, Cimiano P. An introduction to question answering over linked data. In: Proceedings of Reasoning on the Web in the Big Data Era — the 10th International Summer School. Athens, Greece: IEEE, 2014. 100−140

11 Bast H, Haussmann E. More accurate question answering on freebase. In: Proceedings of the 24th International Conference on Information and Knowledge Management. Melbourne, VIC, Australia: ACM, 2015. 1431−1440

12 Bordes A, Chopra S, Weston J. Question answering with subgraph embeddings. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. Doha, Qatar: ACL, 2014. 615−620

13 Yih W, Chang M W, He X D, Gao J F. Semantic parsing via staged query graph generation: Question answering with knowledge base. In: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing. Beijing, China: ACL, 2015. 1321−1331

14 Sun Y W, Zhang L L, Cheng G, Qu Y Z. Sparqa: Skeleton-based semantic parsing for complex questions over knowledge bases. In: Proceedings of the 34th AAAI Conference on Artificial Intelligence. New York, USA: arXiv: 2003.13956, 2020.

15 Xu K, Wu L F, Wang Z G, Yu M, Chen L W, Sheinin V. Exploiting rich syntactic information for semantic parsing with graph-to-sequence model. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Brussels, Belgium: ACL, 2018. 918−924

16 Miller A, Fisch A, Dodge J, Karimi A H, Weston J. Key-value memory networks for directly reading documents. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. Austin, Texas, USA: ACL, 2016. 1400−1409

17 Bordes A, Usunier N, Chopra S, Weston J. Large-scale simple question answering with memory networks. arXiv: 1506.02075. 2015.

18 Yin J, Jiang X, Lu Z D, Shang L F, Li H, Li X M. Neural generative question answering. In: Proceedings of the 25th International Joint Conference on Artificial Intelligence. New York, USA: IJCAI/AAAI, 2016. 2972−2978

19 Liu C, He S Z, Liu K, Zhao J. Curriculum learning for natural answer generation. In: Proceedings of the 27th International Joint Conference on Artificial Intelligence. Stockholm, Sweden: IJCAI/AAAI, 2018. 4223−4229

20 Wang T Z, Cai M, Li J X. A neural conversational model using MMI-WMD decoder based on the Seq2Seq with attention mechanism. In: Proceedings of the 2019 Chinese Control and Decision Conference (CCDC). Nanchang, China: IEEE, 2019. 2696−2700

21 Sharma A, Contractor D, Kumar H, Joshi S. Neural conversational QA: Learning to reason vs exploiting patterns. arXiv: 1909.03759, 2019.

22 Lei W Q, Jin X, Kan M Y, Ren Z C, He X N, Yin D W. Sequicity: Simplifying task-oriented dialogue systems with single sequence-to-sequence architectures. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. Melbourne, Australia: ACL, 2018. 1437−1447

23 Rashkin H, Smith E M, Li M, Boureau Y L. Towards empathetic open-domain conversation models: A new benchmark and dataset. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence, Italy: ACL, 2019. 5370−5381

24 Lin Y K, Liu Z Y, Sun M S. Modeling relation paths for representation learning of knowledge bases. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. Lisbon, Portugal: ACL, 2015. 705−714

25 Quan W, Mao Z D, Wang B, Li G. Knowledge graph embedding: A survey of approaches and applications. *IEEE Transactions on Knowledge and Data Engineering*, 2017, **29**(12): 2724−2743

26 Bordes A, Weston J, Usunier N. Open question answering with weakly supervised embedding models. In: Proceedings of the 2014 Machine Learning and Knowledge Discovery in Databases European Conference. Nancy, France: Springer, 2014. 165−180

27 Bahdanau D, Cho K Y, Bengio Y. Neural machine translation by jointly learning to align and translate. Arxiv: 1409.0473, 2014

28 Feng Chong, Shi Ge, Guo Yu-Hang, Gong Jing, Huang He-Yan. An entity linking method for microblog based on semantic categorization by word embeddings. *Acta Automatica Sinica*, 2016, **42**(6): 915−922
(冯冲, 石戈, 郭宇航, 龚静, 黄河燕. 基于词向量语义分类的微博实体链接方法. 自动化学报, 2016, **42**(6): 915−922)

29 Yin W P, Yu M, Xiang B, Zhou B, Schutze H. Simple question answering by attentive convolutional neural network. ArXiv: 1606.03391, 2016.

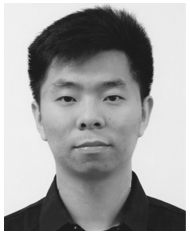
30 Yu M, Yin W P, Hasan K S, Santos C D, Xiang B, Zhou B W. Improved neural relation detection for knowledge base question answering. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. Vancouver, Canada: ACL, 2017. 571−581

31 Deng Y, Xie Y X, Li Y L, Yang M, Shen Y. Multi-task learning with multi-view attention for answer selection and knowledge base question answering. In: Proceedings of the 33rd Conference on Artificial Intelligence. Honolulu, Hawaii, USA: AAAI, 2019. 6318−6325

刘琼昕 北京理工大学计算机学院副教授. 主要研究方向为人工智能, 自然语言处理, 具体研究知识推理, 关系抽取, 任务规划, 决策支持. 本文通信作者. E-mail: summer@bit.edu.cn
(**LIU Qiong-Xin** Associate professor at the School of Computer Sci-

ence and Technology, Beijing Institute of Technology. Her research interest covers artificial intelligence and natural language processing, specifically knowledge reasoning, relationship extraction, task planning, and decision support. Corresponding author of this paper.)

王亚男 北京理工大学计算机学院硕士研究生. 主要研究方向为自然语言处理, 问答系统.
E-mail: wyn1895@163.com
(**WANG Ya-Nan** Master student at the School of Computer Science and Technology, Beijing Institute of Technology. Her research interest covers natural language processing and question answering system.)



龙 航 北京理工大学计算机学院硕士研究生. 主要研究方向为自然语言处理, 表示学习, 问答系统.
E-mail: longhang@ict.ac.cn
(**LONG Hang** Master student at the School of Computer Science and Technology, Beijing Institute of Technology. His research interest covers natural language processing, representation learning, and question answering system.)

王佳升 北京理工大学硕士研究生. 主要研究方向为深度学习, 自然语言处理和知识图谱.
E-mail: 3120191049@bit.edu.cn
(**WANG Jia-Sheng** Master student at the School of Computer Science and Technology, Beijing Institute of Technology. His research interest covers deep learning, natural language processing, and knowledge graphs.)

卢士帅 北京理工大学计算机学院硕士研究生. 主要研究方向为自然语言处理领域的关系提取, 特别是关系提取中的小样本学习.
E-mail: 3120191028@bit.edu.cn
(**LU Shi-Shuai** Master student at the School of Computer Science and Technology, Beijing Institute of Technology. His research interest covers relation extraction in the field of natural language processing, especially few-shot-learning in relation extraction.)